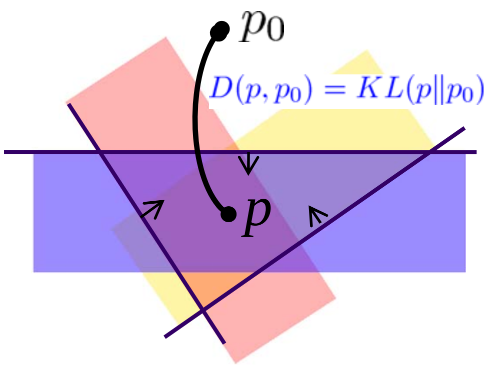
School of Computer Science
Carnegie Mellon

Probabilistic Graphical Models

Posterior Regularization: an integrative paradigm for learning GMs



Eric Xing
(courtesy to Jun Zhu)

Lecture 29, April 30, 2014

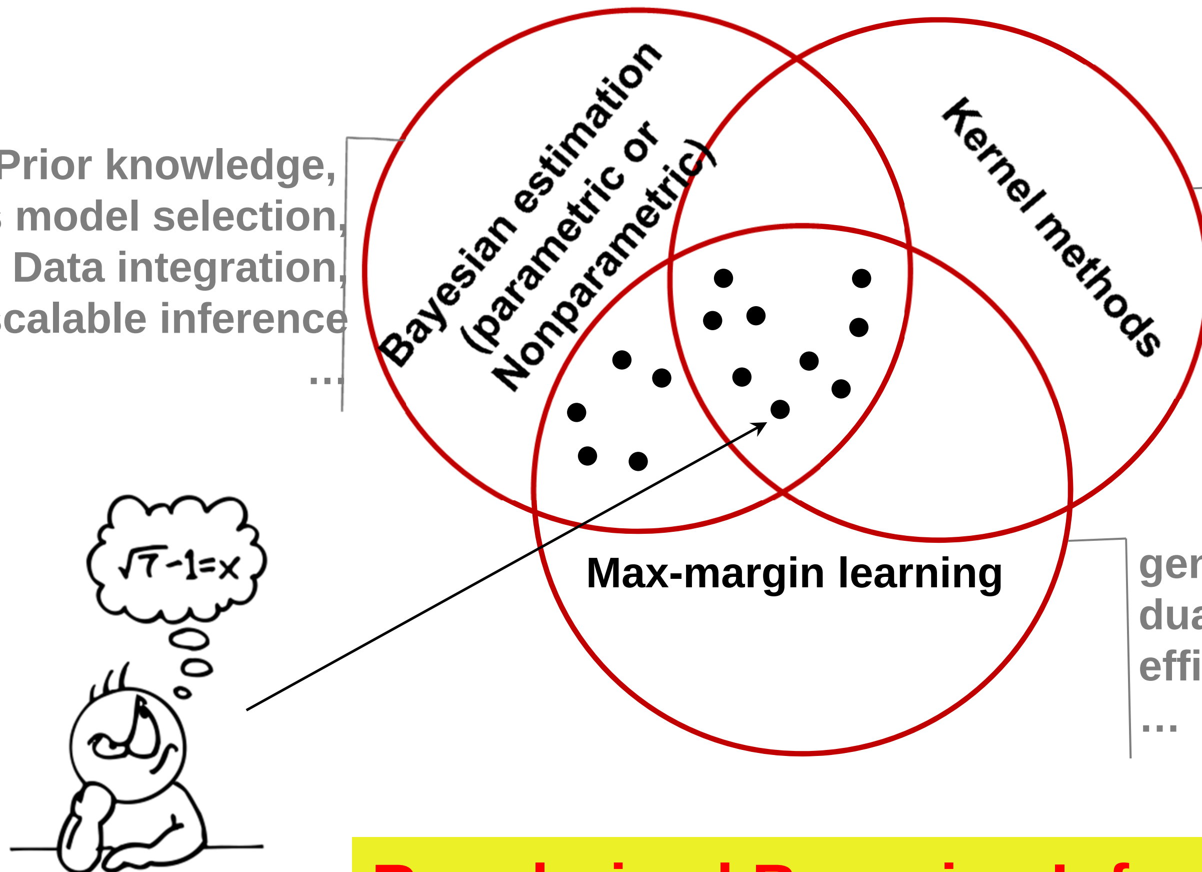
Reading:

Learning GMs

Prior knowledge,
bypass model selection,
Data integration,
scalable inference
…

nonlinear transformation
rich forms of data
…

generalization
dual sparsity
efficient solvers
…



Regularized Bayesian Inference

Bayesian Inference

- A coherent framework of dealing with uncertainties

$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}}$$

- ***M*: a model from some hypothesis space**
- **x: observed data**

Thomas Bayes (1702 – 1761)

- Bayes' rule offers a mathematically rigorous computational mechanism for combining prior knowledge with incoming evidence

Parametric Bayesian Inference

$\mathcal{M}$ **is represented as a finite set of parameters** θ

- A parametric likelihood: $\mathbf{x} \sim p(\cdot|\theta)$
- Prior on θ: $\pi(\theta)$
- Posterior distribution

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\pi(\theta)}{\int p(\mathbf{x}|\theta)\pi(\theta)d\theta} \propto p(\mathbf{x}|\theta)\pi(\theta)$$

Examples:

- **Gaussian distribution prior + 2D Gaussian likelihood → Gaussian posterior distribution**
- **Dirichilet distribution prior + 2D Multinomial likelihood → Dirichlet posterior distribution**
- **Sparsity-inducing priors + some likelihood models → Sparse Bayesian inference**

Nonparametric Bayesian Inference

$\mathcal{M}$ **is a richer model, e.g., with an infinite set of parameters**

- A nonparametric likelihood: $\mathbf{x} \sim p(\cdot|\mathcal{M})$
- Prior on $\mathcal{M}$: $\pi(\mathcal{M})$
- Posterior distribution

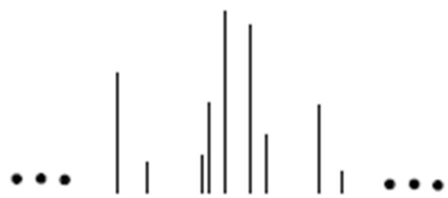
$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}} \propto p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})$$

Examples:

→ see next slide

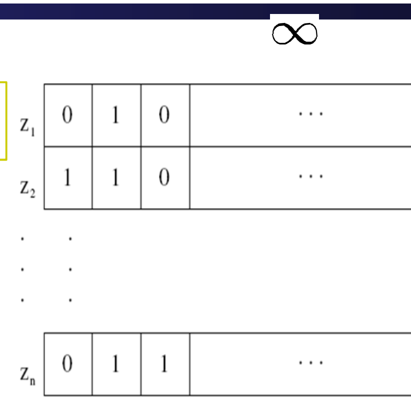
Nonparametric Bayesian Inference

probability measure



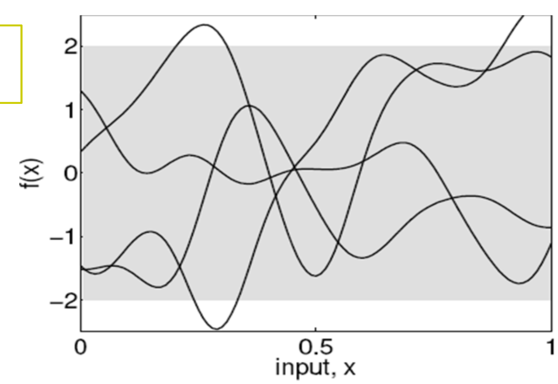
Dirichlet Process Prior [Antoniak, 1974]
+ Multinomial/Gaussian/Softmax likelihood

binary matrix



Indian Buffet Process Prior [Griffiths & Gharamani, 2005]
+ Gaussian/Sigmoid/Softmax likelihood

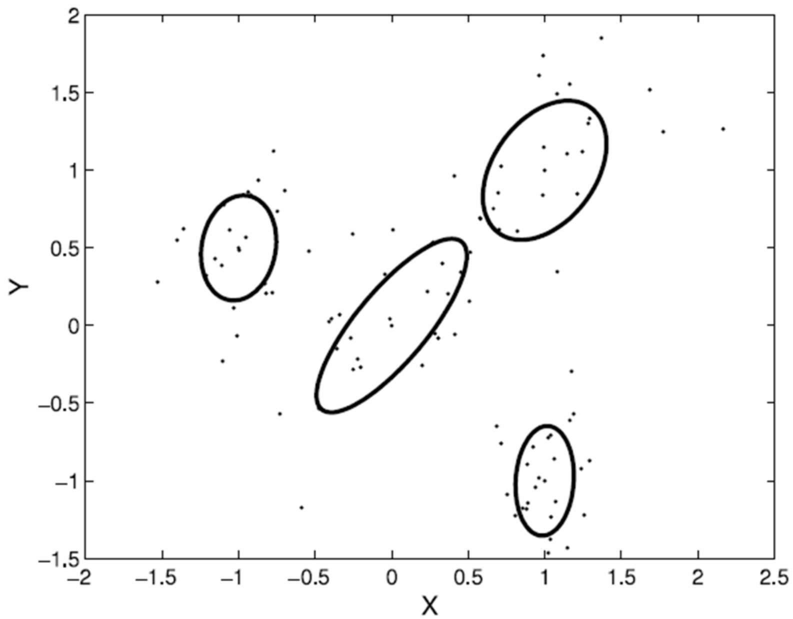
function



Gaussian Process Prior [Doob, 1944; Rasmussen & Williams, 2006]
+ Gaussian/Sigmoid/Softmax likelihood

Why Bayesian Nonparametrics?

- Let the data speak for themselves
- Bypass the model selection problem
  - let data determine model complexity (e.g., the number of components in mixture models)
  - allow model complexity to grow as more data observed



Can we further control the posterior distributions?

posterior **likelihood model** **prior**

$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}}$$

It is desirable to further regularize the posterior distribution

- An extra freedom to perform Bayesian inference
- Arguably more direct to control the behavior of models
- Can be easier and more natural in some examples

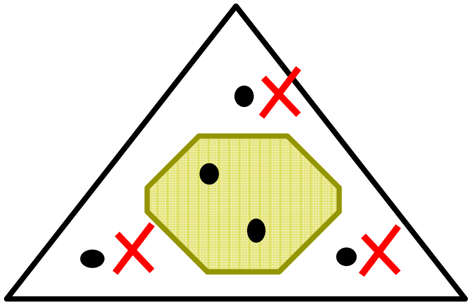
Can we further control the posterior distributions?

posterior — likelihood model — prior

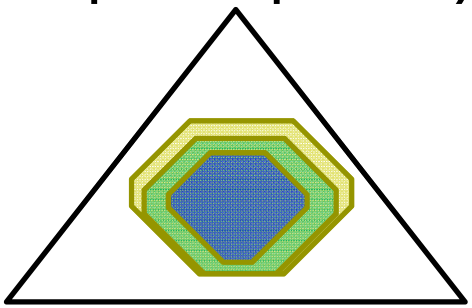
$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}}$$

- **Directly control the posterior distributions?**
  - **Not obvious how …**

hard constraints
(A single feasible space)



soft constraints
(many feasible subspaces with different complexities/penalties)



A reformulation of Bayesian inference

posterior — likelihood model — prior

$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}}$$

- **Bayes' rule is equivalent to:**

$$\begin{aligned} \min_{p(\mathcal{M})} \quad & \mathrm{KL}(p(\mathcal{M})\|\pi(\mathcal{M})) - \mathbb{E}_{p(\mathcal{M})}[\log p(\mathbf{x}|\mathcal{M})] \\ \mathrm{s.t.}: \quad & p(\mathcal{M}) \in \mathcal{P}_{\mathrm{prob}}, \end{aligned}$$

A direct but trivial constraint on the posterior distribution

E.T. Jaynes (1988): "this fresh interpretation of Bayes' theorem could make the use of Bayesian methods more attractive and widespread, and stimulate new developments in the general theory of inference"

[Zellner, Am. Stat. 1988]

Regularized Bayesian Inference

$$\inf_{q(\mathrm{M}),\xi} \mathrm{KL}(q(\mathrm{M})\|\pi(\mathrm{M})) - \int_{\mathcal{M}} \log p(\mathcal{D}|\mathrm{M})q(\mathrm{M})d\mathrm{M} + U(\xi)$$
$$\text{s.t.} : q(\mathrm{M}) \in \mathcal{P}_{\text{post}}(\xi),$$

where, **e.x.**,

$$\mathcal{P}_{\text{post}}(\xi) \stackrel{\text{def}}{=} \Big\{q(\mathrm{M}) |\ \forall t = 1, \cdots, T,\ h\big(Eq(\psi_t; \mathcal{D})\big) \le \xi_t\Big\}.$$

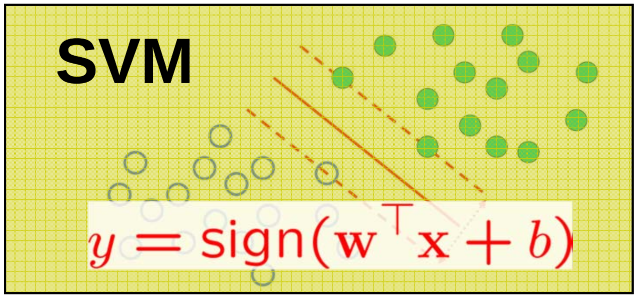
and

$$U(\xi) = \sum_{t=1}^{T} \mathbb{I}(\xi_t = \gamma_t) = \mathbb{I}(\xi = \gamma)$$

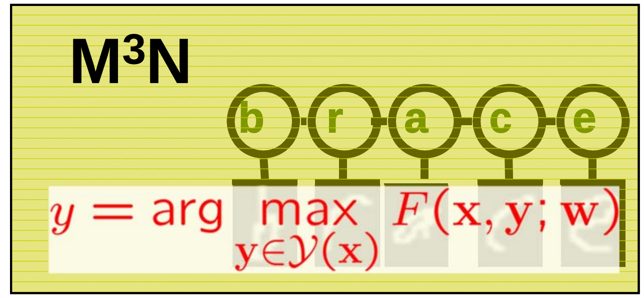
Solving such constrained optimization problem needs convex duality theory

So, where does the constraints come from?

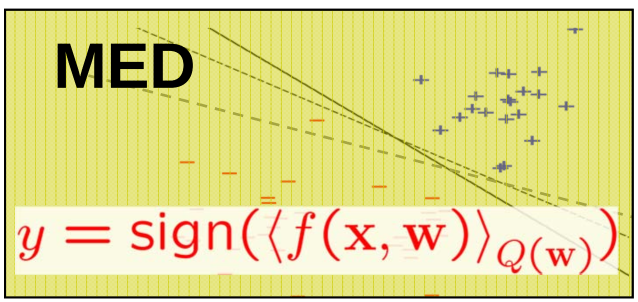
Recall our evolution of the Max-Margin Learning Paradigms



$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m} \xi_i$$
$$y^i(\mathbf{w}^\top \mathbf{x}^i + b) \geq 1 - \xi_i, \quad \forall i$$



$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m} \xi_i$$
$$\mathbf{w}^\top[\mathbf{f}(\mathbf{x}^i,\mathbf{y}^i) - \mathbf{f}(\mathbf{x}^i,\mathbf{y})] \geq \ell(\mathbf{y}^i,\mathbf{y}) - \xi_i, \quad \forall i, \forall \mathbf{y} \neq \mathbf{y}^i$$



$$\min_{Q} \quad \text{KL}(Q||Q_0)$$
$$y^i\langle f(\mathbf{x}^i)\rangle_Q \geq \xi_i, \quad \forall i$$

MED-MN ?

= SMED + "Bayesian" M³N

Maximum Entropy Discrimination Markov Networks

- Structured MaxEnt Discrimination (SMED):

$$\text{P1}: \quad \min_{p(\mathbf{w}),\xi} \boxed{KL(p(\mathbf{w})||p_0(\mathbf{w})) + U(\xi)}$$
$$\text{s.t. } \; p(\mathbf{w}) \in \mathcal{F}_1, \; \xi_i \geq 0, \forall i.$$

generalized maximum entropy or *regularized* KL-divergence

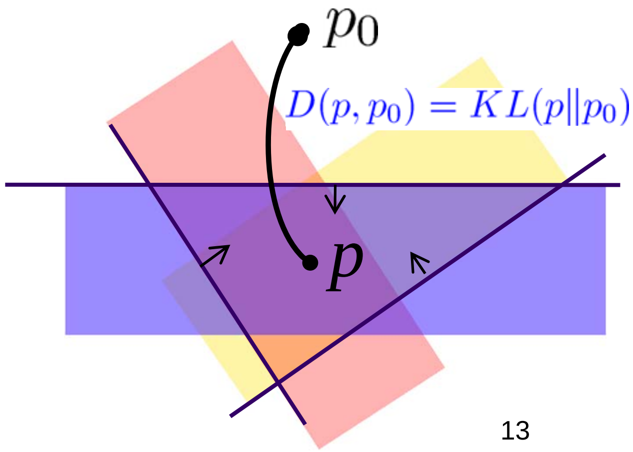
- Feasible subspace of weight distribution:

$$\mathcal{F}_1 = \{p(\mathbf{w}) : \boxed{\int p(\mathbf{w})[\Delta F_i(\mathbf{y};\mathbf{w}) - \Delta\ell_i(\mathbf{y})]\, d\mathbf{w} \geq -\xi_i,} \; \forall i, \forall \mathbf{y} \neq \mathbf{y}^i\},$$

expected margin constraints.

- Average from distribution of M³Ns

$$h_1(\mathbf{x}; p(\mathbf{w})) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \int p(\mathbf{w}) F(\mathbf{x},\mathbf{y};\mathbf{w}) d\mathbf{w}$$



Can we use this scheme to learn models other than MN?

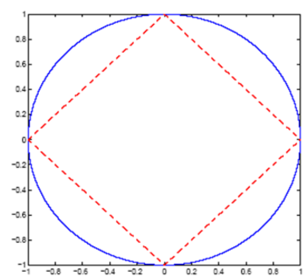
Recall the 3 advantages of MEDN

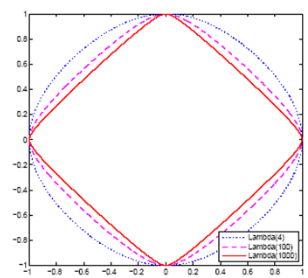
- An averaging Model: PAC-Bayesian prediction error guarantee (Theorem 3)

$$\Pr_Q(M(h,\mathbf{x},\mathbf{y}) \le 0) \le \Pr_{\mathcal{D}}(M(h,\mathbf{x},\mathbf{y}) \le \gamma) + O\Big(\sqrt{\frac{\gamma^{-2}KL(p\|p_0)\ln(N|\mathcal{Y}|) + \ln N + \ln \delta^{-1}}{N}}\Big).$$

- Entropy regularization: Introducing useful biases
  - Standard Normal prior => reduction to standard M³N (we've seen it)
  - Laplace prior => Posterior shrinkage effects (sparse M³N)

$$\min_{\mu,\xi} \ \sqrt{\lambda}\sum_{k=1}^{K}\Big(\sqrt{\mu_k^2 + \frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}}\log\frac{\sqrt{\lambda\mu_k^2+1}+1}{2}\Big) + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t.} \ \ \mu^\top \Delta\mathbf{f}_i(\mathbf{y}) \ge \Delta\ell_i(\mathbf{y}) - \xi_i; \ \xi_i \ge 0, \ \ \forall i, \ \forall \mathbf{y} \neq \mathbf{y}^i.$$

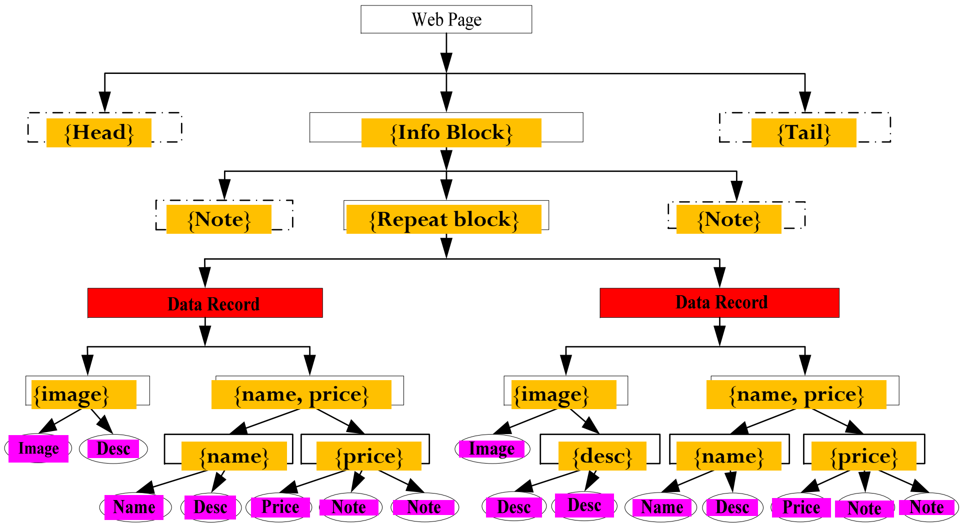




- Integrating Generative and Discriminative principles (next class)
  - Incorporate latent variables and structures (PoMEN)
  - Semisupervised learning (with partially labeled data)

Latent Hierarchical MaxEnDNet

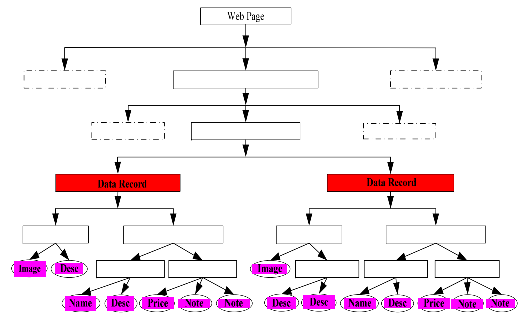
- Web data extraction
  - Goal: *Name*, *Image*, *Price*, *Description*, etc.



- **Hierarchical labeling**
- **Advantages:**
  - Computational efficiency
  - Long-range dependency
  - Joint extraction

Partially Observed MaxEnDNet (PoMEN) **(Zhu et al, NIPS 2008)**

- Now we are given partially labeled data: $\mathcal{D} = \{< \mathbf{x}^i, \mathbf{y}^i, \mathbf{z}^i >\}_{i=1}^N$



  - PoMEN: learning $p(\mathbf{w}, \mathbf{z})$

$$\mathrm{P2(PoMEN)}: \quad \min_{p(\mathbf{w},\{\mathbf{z}\}),\xi} KL(p(\mathbf{w},\{\mathbf{z}\})||p_0(\mathbf{w},\{\mathbf{z}\})) + U(\xi)$$
$$\text{s.t. } p(\mathbf{w},\{\mathbf{z}\}) \in \mathcal{F}_2, \ \xi_i \geq 0, \forall i.$$

$$\mathcal{F}_2 = \{p(\mathbf{w},\{\mathbf{z}\}) : \sum_{\mathbf{z}} \int p(\mathbf{w},\mathbf{z})[\Delta F_i(\mathbf{y},\mathbf{z};\mathbf{w}) - \Delta\ell_i(\mathbf{y})]\,\mathrm{d}\mathbf{w} \geq -\xi_i, \ \forall i, \forall \mathbf{y} \neq \mathbf{y}^i\},$$

  - Prediction: $h_2(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \sum_{\mathbf{z}} \int p(\mathbf{w},\mathbf{z}) F(\mathbf{x},\mathbf{y},\mathbf{z};\mathbf{w})\,\mathrm{d}\mathbf{w}$

Alternating Minimization Alg.

- Factorization assumption:

$$p_0(\mathbf{w}, \{\mathbf{z}\}) = p_0(\mathbf{w}) \prod_{i=1}^{N} p_0(\mathbf{z}_i) \qquad\qquad p(\mathbf{w}, \{\mathbf{z}\}) = p(\mathbf{w}) \prod_{i=1}^{N} p(\mathbf{z}_i)$$

- Alternating minimization:
  - Step 1: keep $p(\mathbf{z})$ fixed, optimize over $p(\mathbf{w})$

$$\min_{p(\mathbf{w}),\xi} \; KL(p(\mathbf{w})||p_0(\mathbf{w})) + C\sum_i \xi_i$$
$$\text{s.t.} \;\; p(\mathbf{w}) \in \mathcal{F}_1', \;\; \xi_i \geq 0, \forall i.$$

$$\mathcal{F}_1' = \{p(\mathbf{w}) : \;\; \int p(\mathbf{w}) E_{p(\mathbf{z})}[\Delta F_i(\mathbf{y}, \mathbf{z}; \mathbf{w}) - \Delta \ell_i(\mathbf{y})] \, \mathrm{d}\mathbf{w} \geq -\xi_i, \; \forall i, \; \forall \mathbf{y}\}$$

o Normal prior
  - M³N problem (QP)

o Laplace prior
  - Laplace M³N problem (VB)

  - Step 2: keep $p(\mathbf{w})$ fixed, optimize over $p(\mathbf{z})$

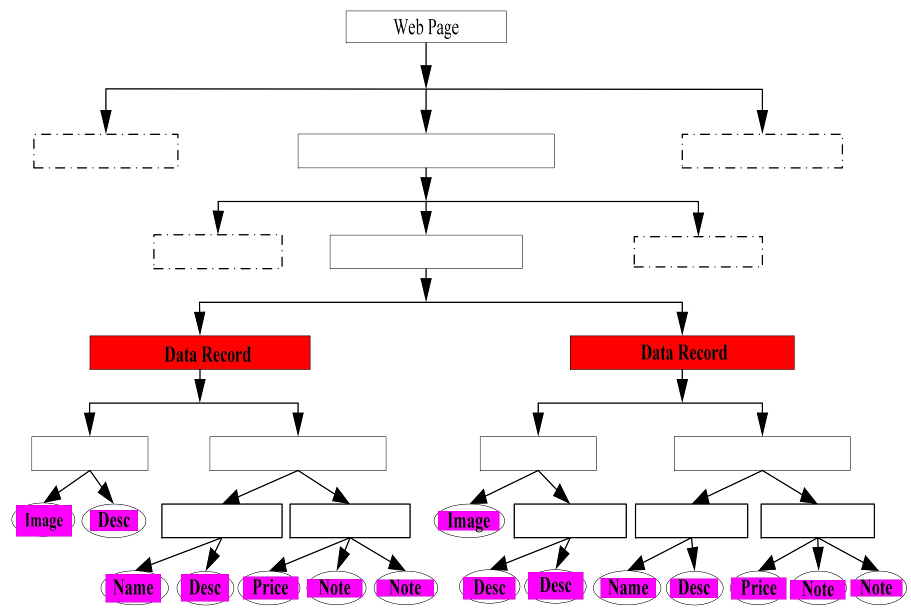
$$\min_{p(\mathbf{w}),\xi} \; KL(p(\mathbf{z})||p_0(\mathbf{z})) + C\xi_i$$
$$\text{s.t.} \;\; p(\mathbf{z}) \in \mathcal{F}_1^\star, \;\; \xi_i \geq 0.$$

$$\mathcal{F}_1^\star = \{p(\mathbf{z}) : \;\; \sum_{\mathbf{z}} p(\mathbf{z}) \int p(\mathbf{w})[\Delta F_i(\mathbf{y}, \mathbf{z}; \mathbf{w}) - \Delta \ell_i(\mathbf{y})] \, \mathrm{d}\mathbf{w} \geq -\xi_i, \; \forall i, \; \forall \mathbf{y}\}$$

Equivalently reduced to an LP with a polynomial number of constraints

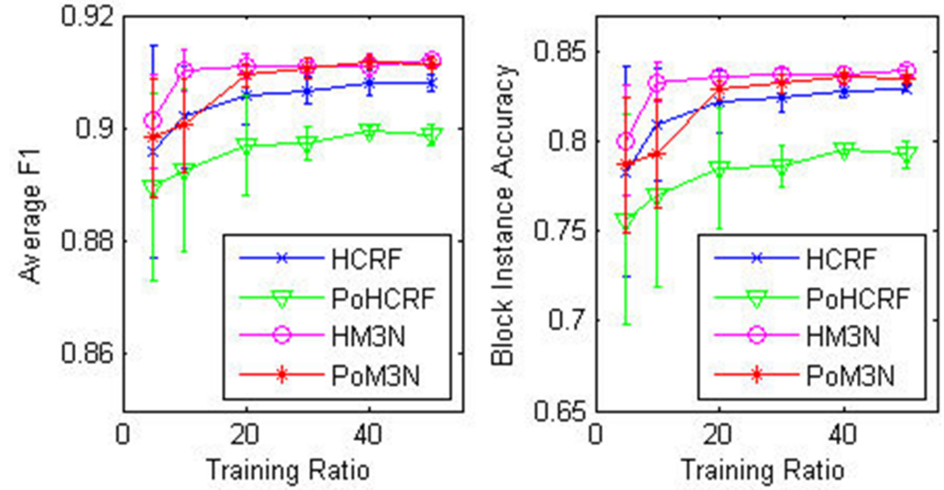
Experimental Results

- Web data extraction:
  - *Name, Image, Price, Description*
  - Methods:
    - Hierarchical CRFs, Hierarchical M^3N
    - PoMEN, Partially observed HCRFs
  - Pages from 37 templates
    - Training: 185 (5/per template) pages, or 1585 data records
    - Testing: 370 (10/per template) pages, or 3391 data records
  - Record-level Evaluation
    - Leaf nodes are labeled
  - Page-level Evaluation
    - Supervision Level 1:
      - Leaf nodes and data record nodes are labeled
    - Supervision Level 2:
      - Level 1 + the nodes above data record nodes

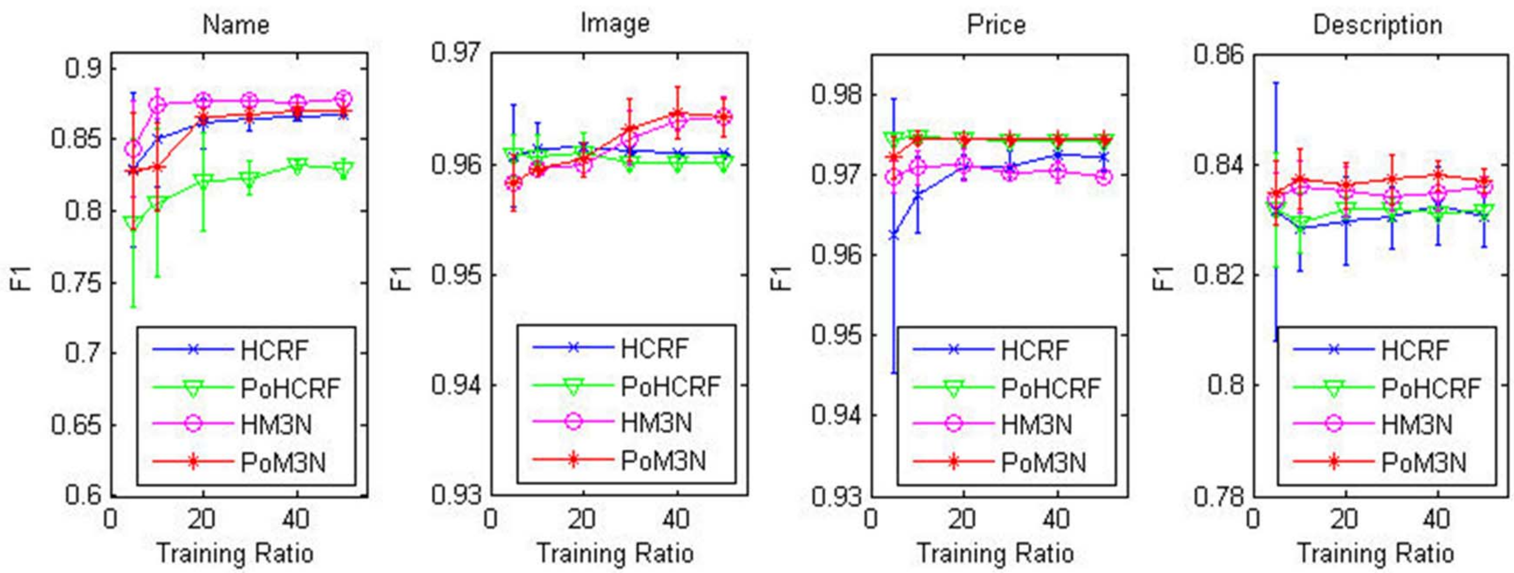


Record-Level Evaluations

- Overall performance:
  - Avg F1:
    - avg F1 over all attributes
  - Block instance accuracy:
    - % of records whose *Name*, *Image*, and *Price* are correct

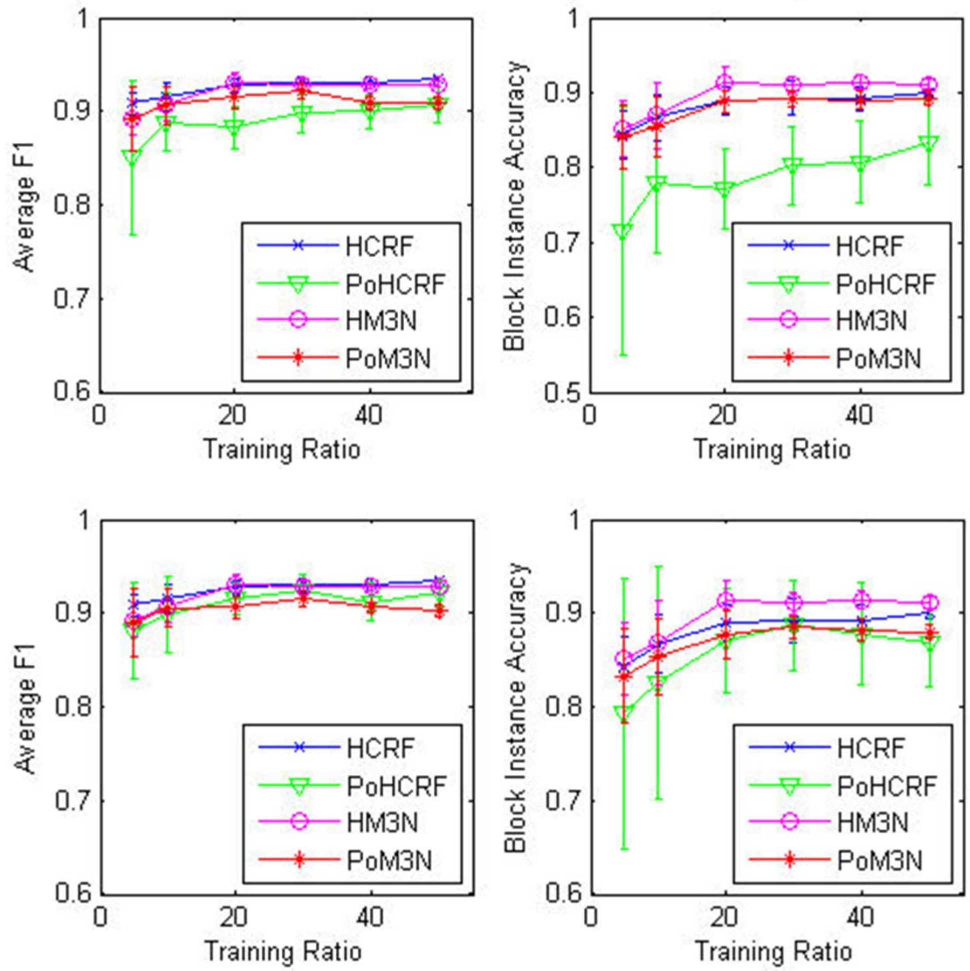


- Attribute performance:



Page-Level Evaluations

- Supervision Level 1:
  - Leaf nodes and data record nodes are labeled
- Supervision Level 2:
  - Level 1 + the nodes above data record nodes



Key message from PoMEN

- Structured MaxEnt Discrimination (SMED):

$$\mathrm{P1}: \quad \min_{p(\mathbf{w},\mathbf{z}),\xi} \boxed{KL(p(\mathbf{w},\mathbf{z})||p_0(\mathbf{w},\mathbf{z})) + U(\xi)}$$
$$, \quad \text{s.t.} \;\; p(\mathbf{w},\mathbf{z}) \in \mathcal{F}_1, \;\; \xi_i \geq 0, \forall i.$$

generalized maximum entropy or *regularized* KL-divergence

- Feasible subspace of weight distribution:

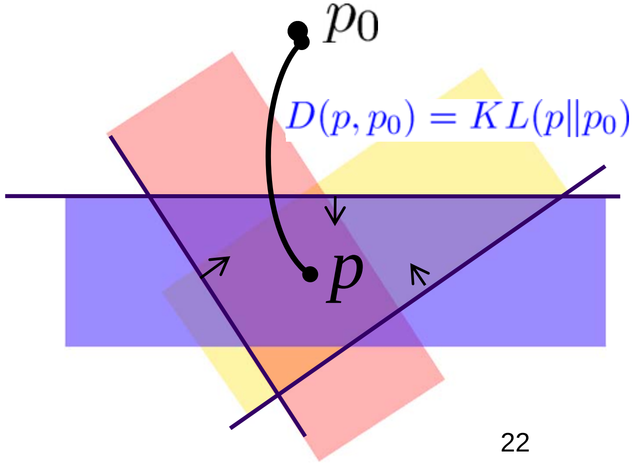
$$\mathcal{F} = \{p(\mathbf{w},\mathbf{z}) : \boxed{\int\int p(\mathbf{w},\mathbf{z})[\Delta F_i(\mathbf{y};\mathbf{w},\mathbf{z}) - \Delta\ell_i(\mathbf{y})]\, \mathrm{d}\mathbf{w}\mathrm{d}\mathbf{z}} \geq -\xi_i, \;\; \forall i, \forall \mathbf{y} \neq \mathbf{y}^i\},$$

expected margin constraints.

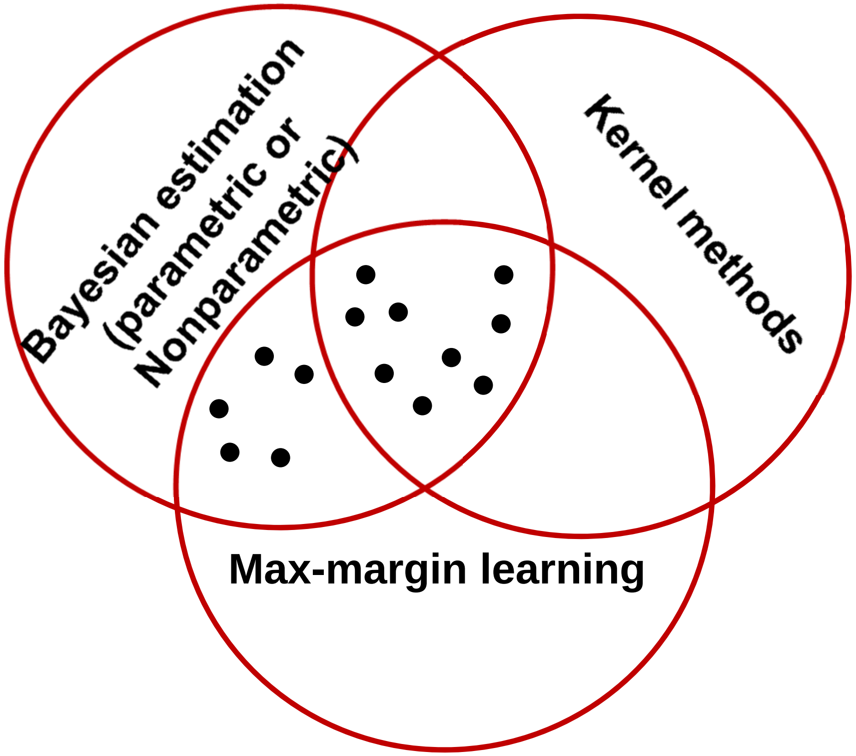
- Average from distribution of PoMENs

$$h(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \int\int p(\mathbf{w},\mathbf{z}) F(\mathbf{x},\mathbf{y},\mathbf{z};\mathbf{w})\, \mathrm{d}\mathbf{w}\mathrm{d}\mathbf{z}$$

- **We can use this for any p and p_0 !**



An all inclusive paradigm for learning general GM --- RegBayes



$$\inf_{q(\mathrm{M}),\xi} \mathrm{KL}(q(\mathrm{M})\|\pi(\mathrm{M})) - \int_{\mathcal{M}} \log p(\mathcal{D}|\mathrm{M})q(\mathrm{M})d\mathrm{M} + U(\xi)$$
$$\text{s.t.}: q(\mathrm{M}) \in \mathcal{P}_{\mathrm{post}}(\xi),$$

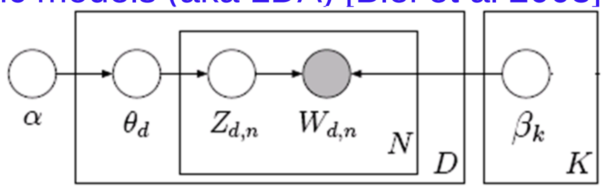
Predictive Latent Subspace Learning

via a large-margin approach

… where M is any subspace model and p is a parametric Bayesian prior

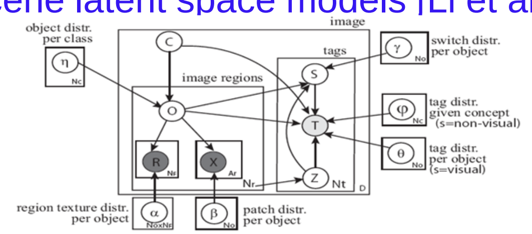
Unsupervised Latent Subspace Discovery

- Finding latent subspace representations (an old topic)
  - Mapping a high-dimensional representation into a latent low-dimensional representation, where each dimension can have some interpretable meaning, e.g., a semantic topic
- Examples:
  - Topic models (aka LDA) [Blei et al 2003]

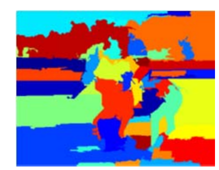


⇒

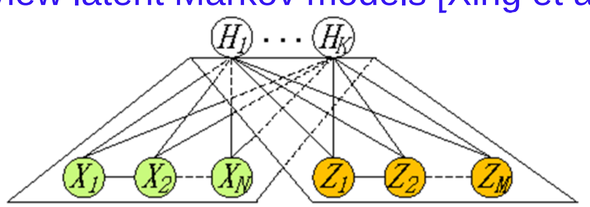
  - Total scene latent space models [Li et al 2009]



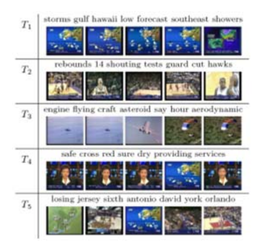
⇒



  - Multi-view latent Markov models [Xing et al 2005]



⇒



  - PCA, CCA, …

Predictive Subspace Learning with *Supervision*

- Unsupervised latent subspace representations are generic but can be sub-optimal for predictions
- Many datasets are available with supervised side information
  - Tripadvisor Hotel Review (http://www.tripadvisor.com)

"Lovely welcomming staff, good rooms that give a good nights sleep, downtown location"
Meramees Hostel

Save Review

SheikhSahib 10 contributions
London

Jul 7, 2009 | Trip type: Friends getaway

This hotel is just of the side streets of Talat Harb, one of the main arteries to downtown Cairo. It is walking distance to the Nile, riverfront hotels, Egyptian Museum, and there are many eateries in the area at night when it is still bustling. Only a short cab ride away from the Old Fatimid Cairo.

The staff are young and very friendly and able to sort out things like mobile chargers, internet, and they have skype installed on their computers which is brilliant. The rooms are nicer then the Luna (nearby) and much quieter as well.

My ratings for this hotel

Value
Rooms
Location
Cleanliness
Service

Date of stay February 2009

Visit was for Leisure

Traveled with With Friends

Member since July 03, 2009

Would you recommend this hotel to a friend? Yes

  - LabelMe http://labelme.csail.mit.edu/

woman entering shop
man walking towards camera
balcony rail
building
sidewalk
manhole
light
person woman walking
ice cream
head
arm
torso

  - Many others

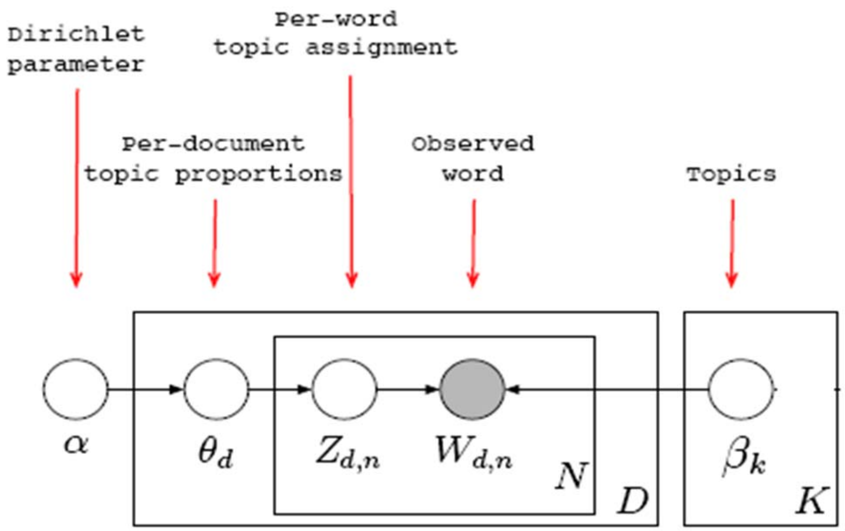
Flickr (http://www.flickr.com/)

IMAGENET

- Can be noisy, but not random noise (Ames & Naaman, 2007)
  - labels & rating scores are usually assigned based on some intrinsic property of the data
  - helpful to suppress noise and capture the most useful aspects of the data
- **Goals:**
  - **Discover latent subspace representations that are both *predictive* and *interpretable* by exploring weak supervision information**

I. LDA: Latent Dirichlet Allocation

(Blei et al., 2003)



- **Generative Procedure:**
  - **For each document *d*:**
    - **Sample a topic proportion** $\theta_d \sim \mathrm{Dir}(\alpha)$
    - **For each word:**
      - **Sample a topic** $Z_{d,n} \sim \mathrm{Mult}(\theta_d)$
      - **Sample a word** $W_{d,n} \sim \mathrm{Mult}(\beta_{z_{d,n}})$

- Joint Distribution: $p(\theta, \mathbf{z}, \mathbf{W}|\alpha, \beta) = \prod_{d=1}^{D} p(\theta_d|\alpha)(\prod_{n=1}^{N} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta))$

  exact inference intractable!

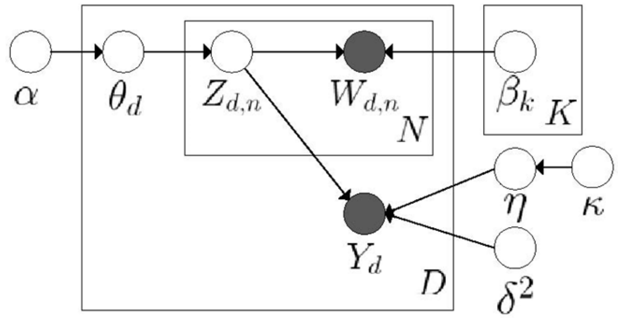
- Variational Inference with $q(\mathbf{z}, \theta) \sim p(\mathbf{z}, \theta|\mathbf{W}, \alpha, \beta)$

$$\mathcal{L}(q) \triangleq -E_q[\log p(\theta, \mathbf{z}, \mathbf{W}|\alpha, \beta)] - \mathcal{H}(q(\mathbf{z}, \theta)) \geq -\log p(\mathbf{W}|\alpha, \beta)$$

- Minimize the variational bound to estimate parameters and infer the posterior distribution

Maximum Entropy Discrimination LDA (MedLDA) (Zhu et al, ICML 2009)

- Bayesian sLDA:



- MED Estimation:
  - MedLDA Regression Model

$$\mathrm{P1}(\mathrm{MedLDA}^r): \quad \min_{q,\alpha,\beta,\delta^2,\xi,\xi^\star} \mathcal{L}(q) + C\sum_{d=1}^{D}(\xi_d + \xi_d^\star)$$

$$\text{s.t. } \forall d: \quad \begin{cases} y_d - E[\eta^\top \bar{Z}_d] \leq \epsilon + \xi_d, & \mu_d \\ -y_d + E[\eta^\top \bar{Z}_d] \leq \epsilon + \xi_d^\star, & \mu_d^\star \\ \xi_d \geq 0, & v_d \\ \xi_d^\star \geq 0, & v_d^\star \end{cases}$$

model fitting

predictive accuracy

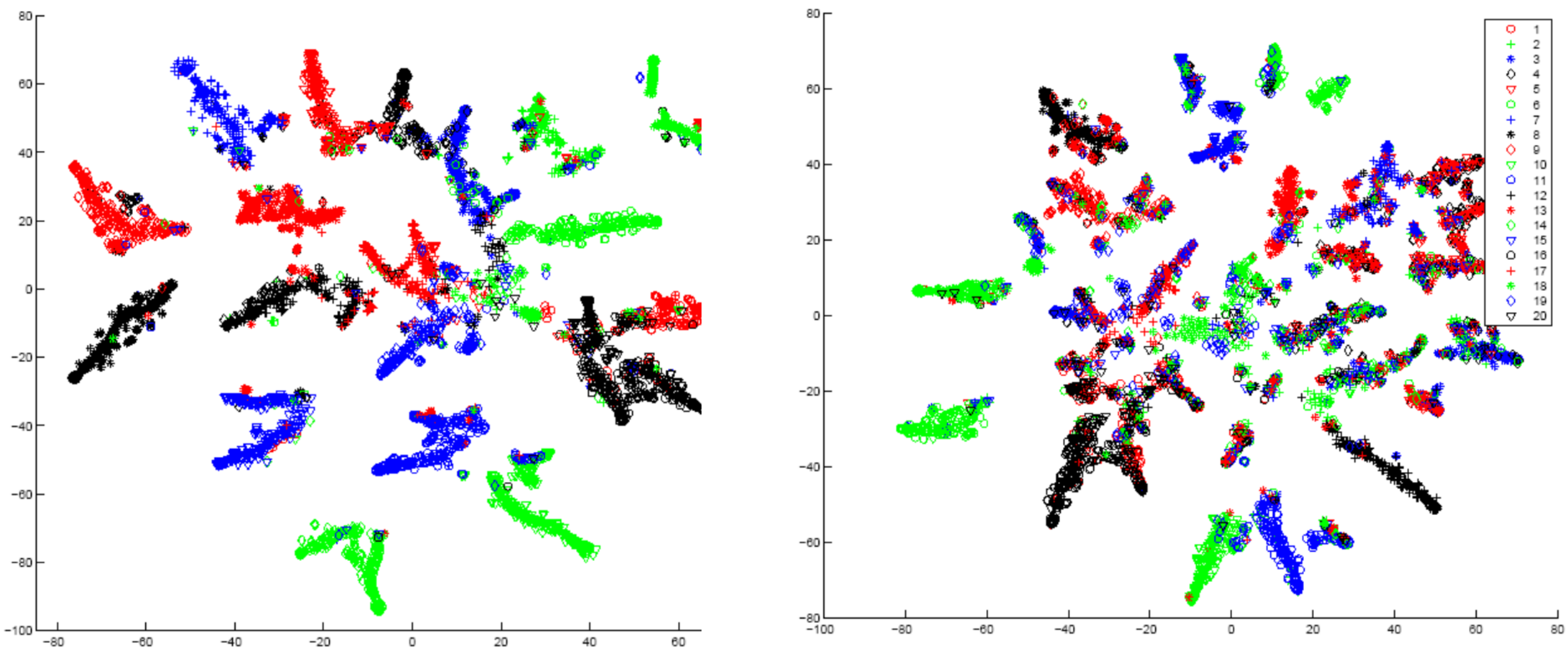
  - MedLDA Classification Model

$$\mathrm{P2}(\mathrm{MedLDA}^c): \quad \min_{q,q(\eta),\alpha,\beta,\xi} \mathcal{L}(q) + C\sum_{d=1}^{D}\xi_d$$

$$\text{s.t. } \forall d,\ y \neq y_d: \quad E[\eta^\top \Delta\mathbf{f}_d(y)] \geq 1 - \xi_d;\ \xi_d \geq 0.$$

Document Modeling

- Data Set: 20 Newsgroups
- 110 topics + 2D embedding with t-SNE (var der Maaten & Hinton, 2008)

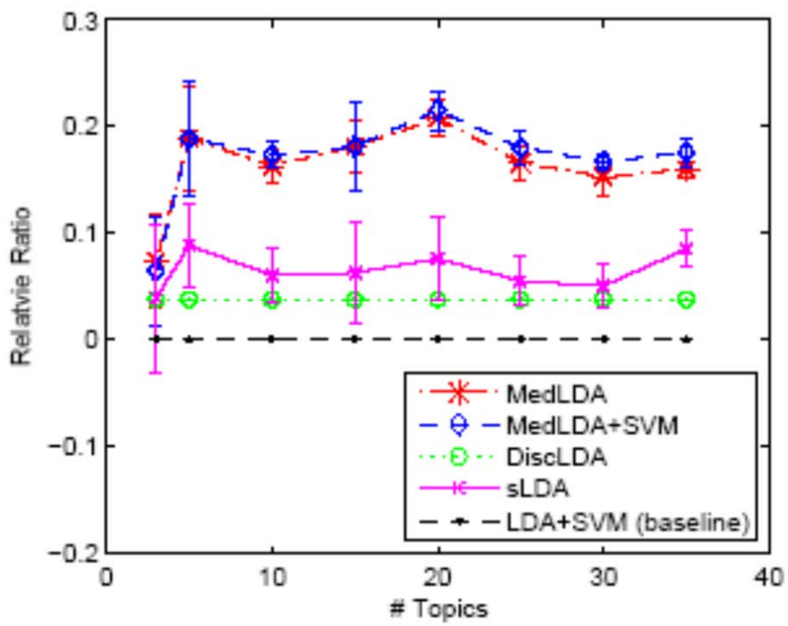


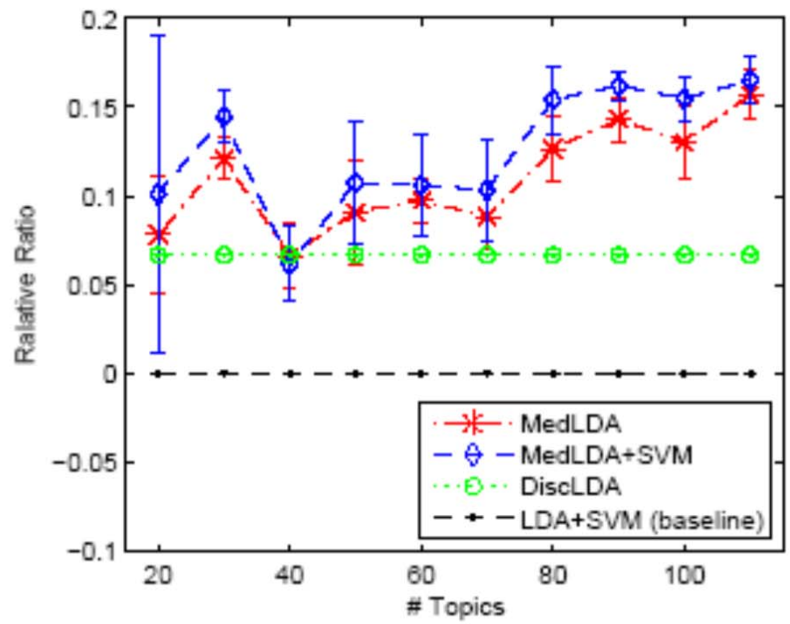
MedLDA **LDA**

Classification

- **Data Set:** 20Newsgroups
  - Binary classification: "alt.atheism" and "talk.religion.misc" (Simon et al., 2008)
  - Multiclass Classification: all the 20 categories
- **Models**: DiscLDA, sLDA (Binary ONLY! Classification sLDA (Wang et al., 2009)), LDA+SVM (baseline), MedLDA, MedLDA+SVM
- **Measure**: Relative Improvement Ratio

$$RR(\mathcal{M}) = \frac{precision(\mathcal{M})}{precision(LDA+SVM)} - 1$$

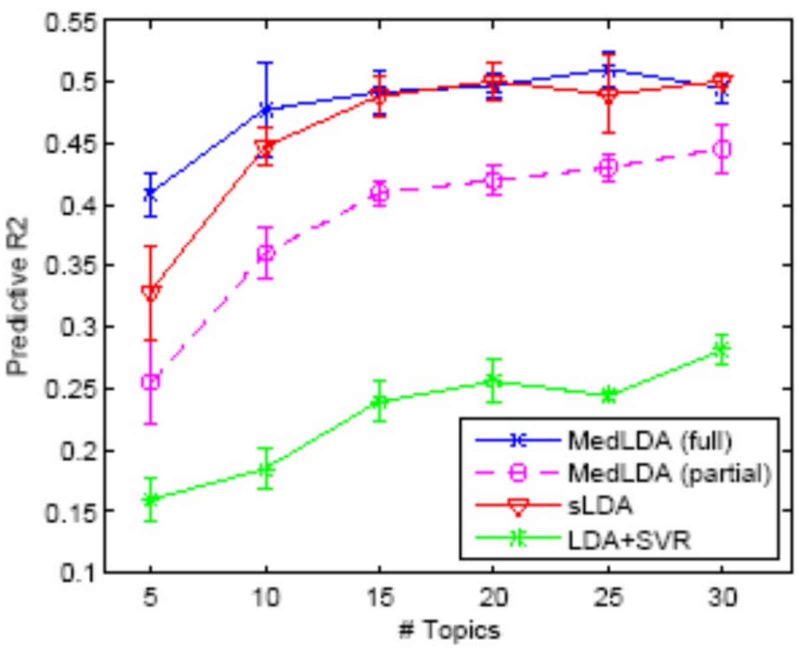


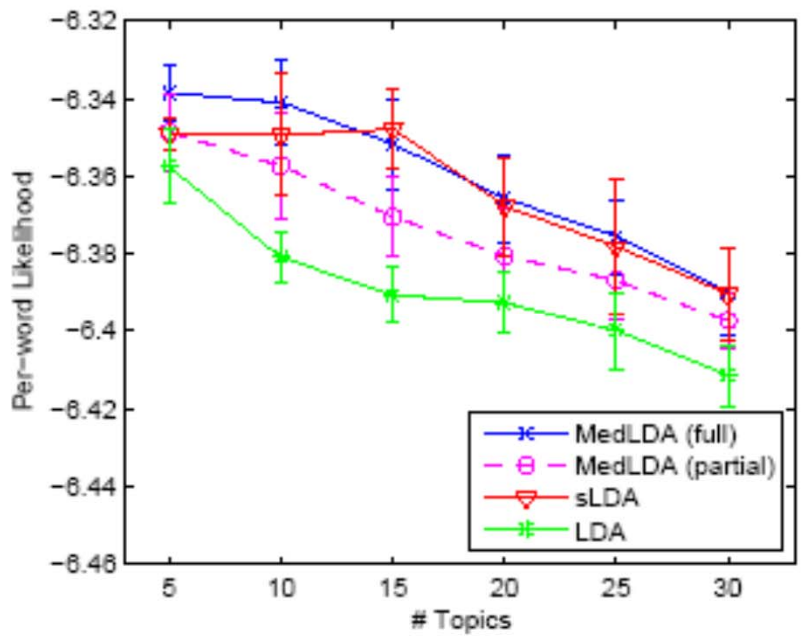


Regression

- **Data Set**: Movie Review (Blei & McAuliffe, 2007)
- **Models**: MedLDA(*partial*), MedLDA(*full*), sLDA, LDA+SVR
- **Measure**: predictive R^2 and per-word log-likelihood

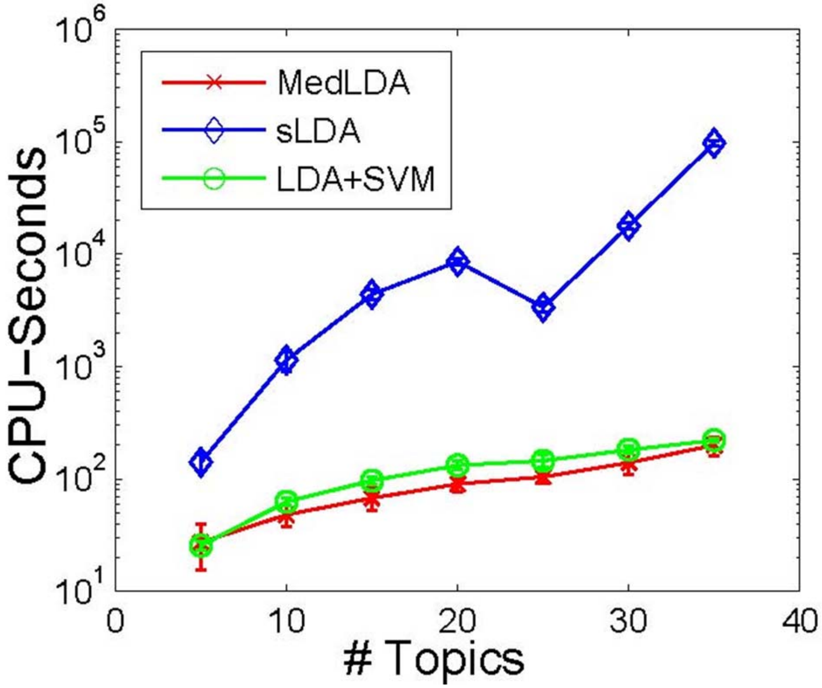
$$pR^2 = 1 - \frac{\sum_d (y_d - \hat{y}_d)^2}{\sum_d (y_d - \bar{y}_d)^2}$$





Time Efficiency

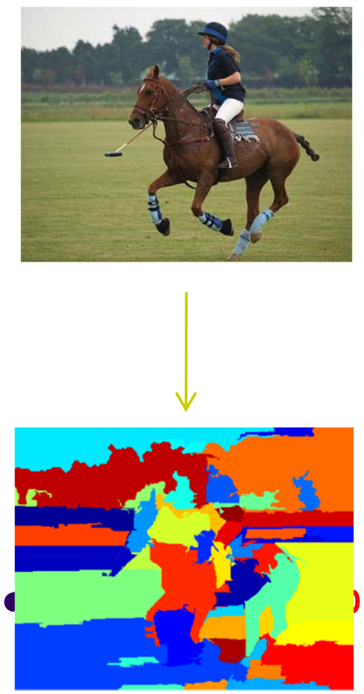
- Binary Classification



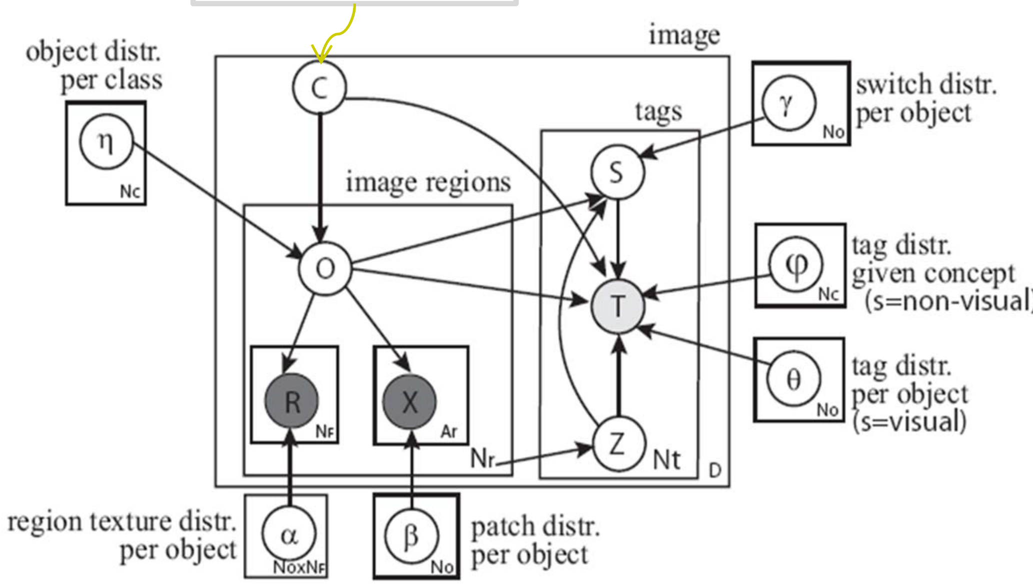
- Multiclass:
  - MedLDA is comparable with LDA+SVM
- Regression:
  - MedLDA is comparable with sLDA

II. Upstream Scene Understanding Models

- The “Total Scene Understanding” Model (Li et al, CVPR 2009)



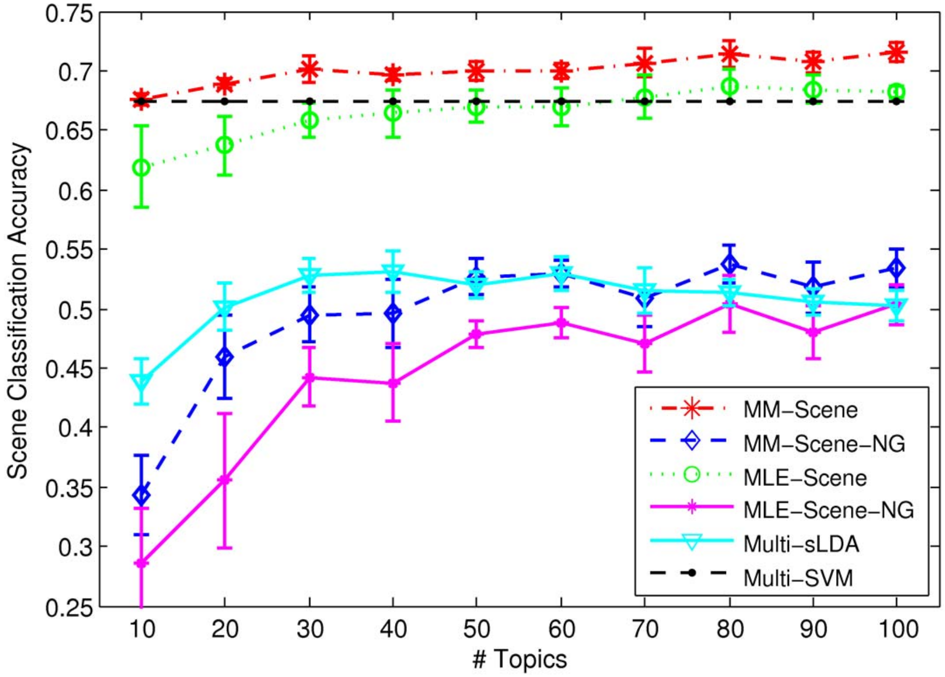
class: Polo



Athlete
Horse
Grass
Trees
Sky
Saddle

Scene Classification

- 8-category sports data set (Li & Fei-Fei, 2007):



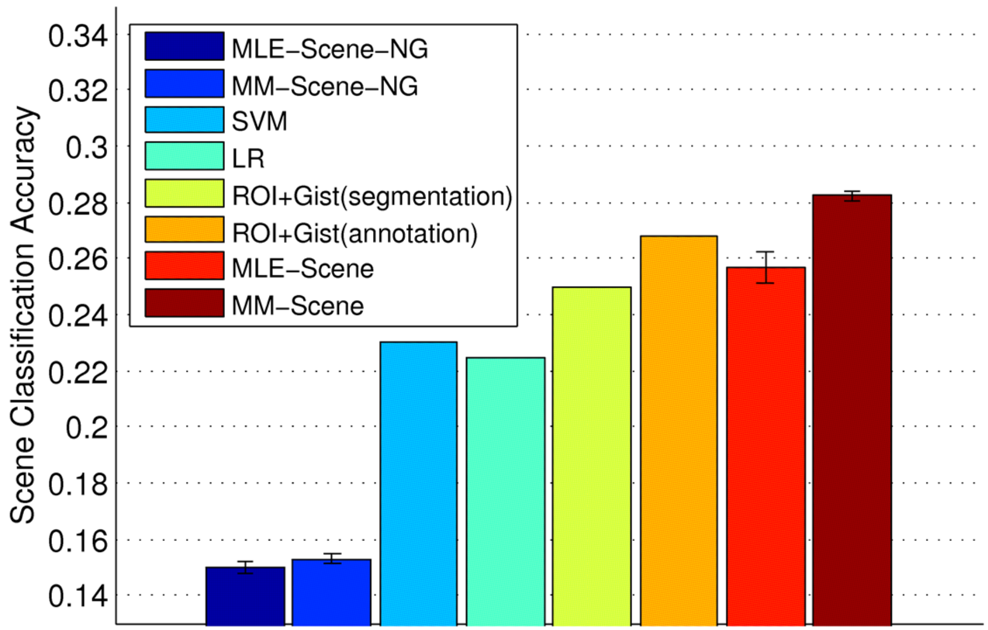
- Fei-Fei's theme model: 0.65 (different image representation)
- SVM: 0.673

•1574 images (50/50 split)
•Pre-segment each image into regions
•Region features:
  •color, texture, and location
  •patches with SIFT features
•Global features:
  •Gist (Oliva & Torralba, 2001)
  •Sparse SIFT codes (Yang et al, 2009)

MIT Indoor Scene

- Classification results:

- **67-category MIT indoor scene (Quattoni & Torralba, 2009):**
  - **~80 per-category for training; ~20 per-category for testing**
  - **Same feature representation as above**
  - **Gist global features**



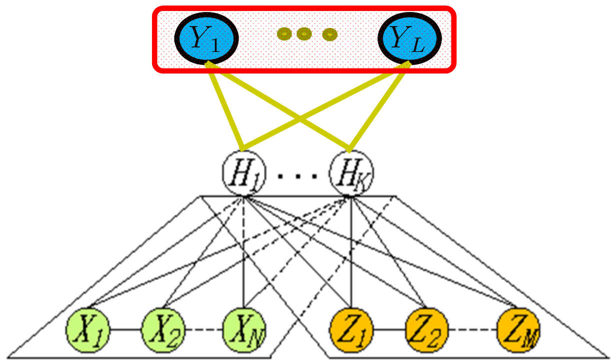
$ROI+Gist(annotation) used *human annotated* interest regions.

III. Supervised Multi-view RBMs

- A probabilistic method with an additional view of response variables *Y*

$$p(y|\mathbf{h}) = \frac{\exp\{\mathbf{V}^\top \mathbf{f}(\mathbf{h}, y)\}}{Z(V, \mathbf{h})}$$

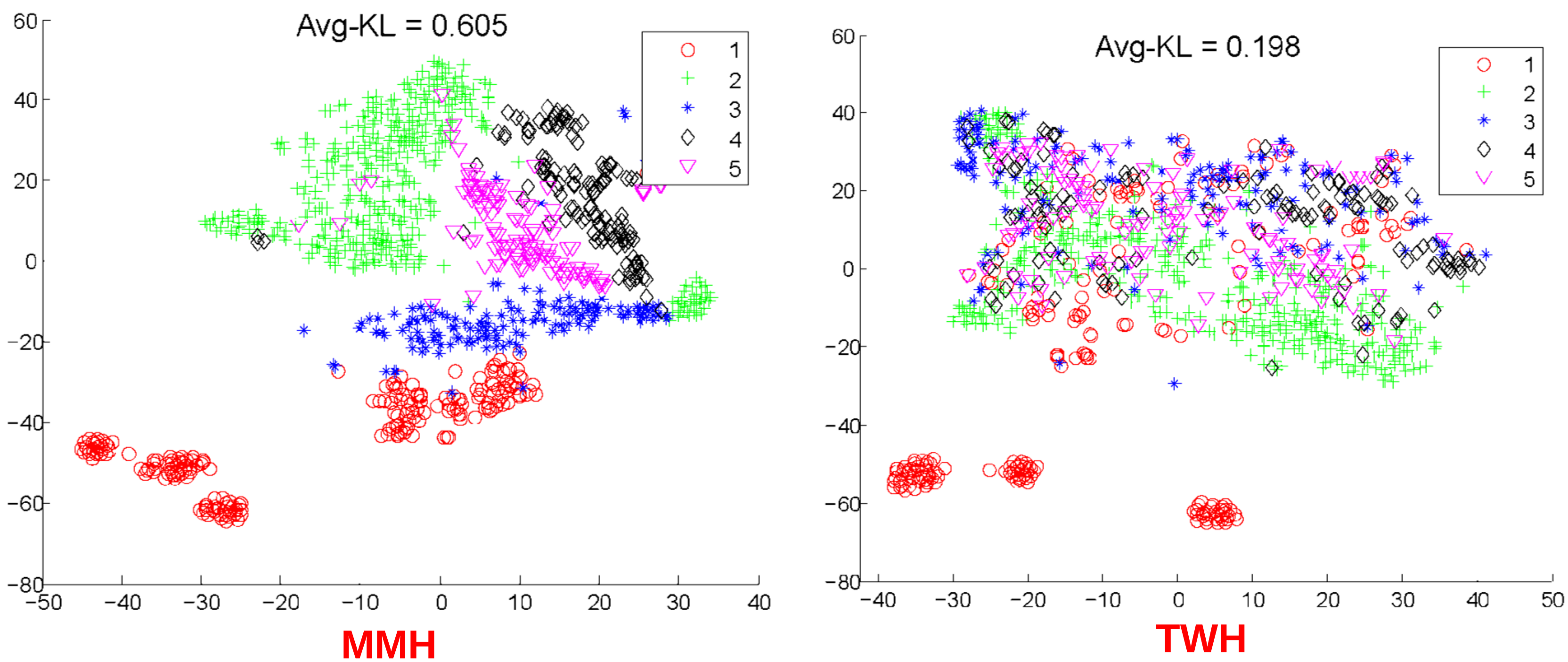
normalization factor



- Parameters can be learned with maximum likelihood estimation, e.g., special supervised Harmonium (Yang et al., 2007)
  - contrastive divergence is the commonly used approximation method in learning undirected latent variable models (Welling et al., 2004; Salakhutdinov & Murray, 2008).

Predictive Latent Representation

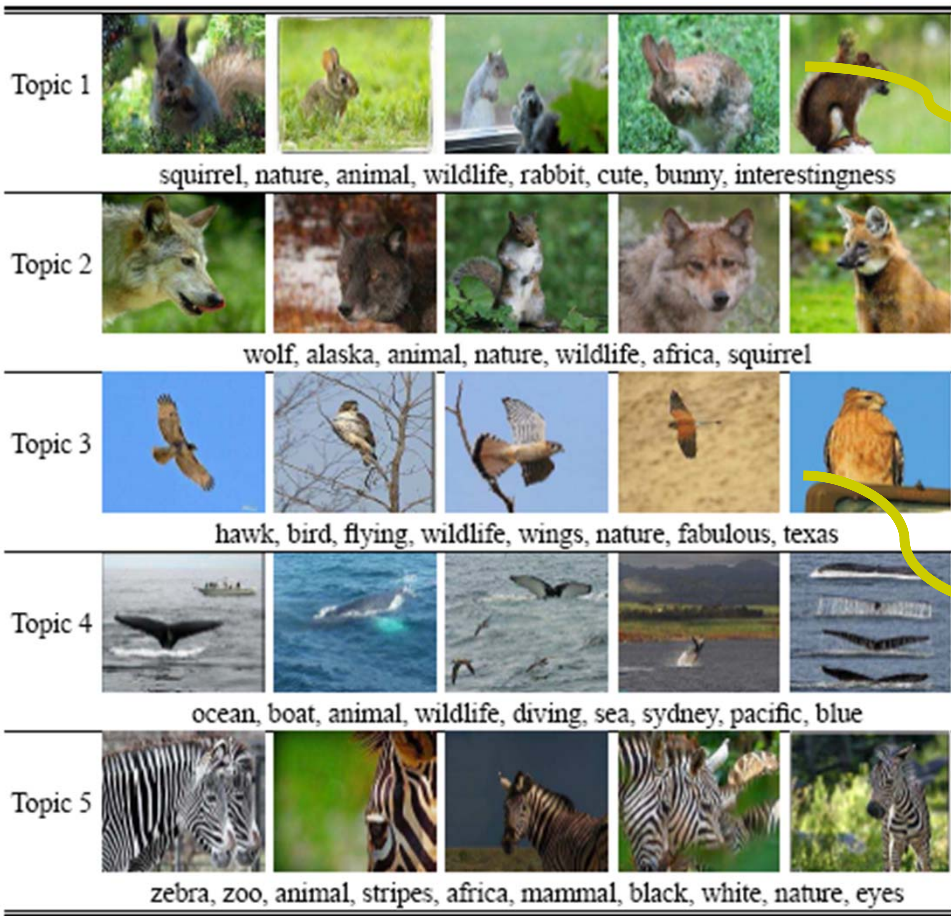
- t-SNE (van der Maaten & Hinton, 2008) 2D embedding of the discovered latent space representation on the TRECVID 2003 data

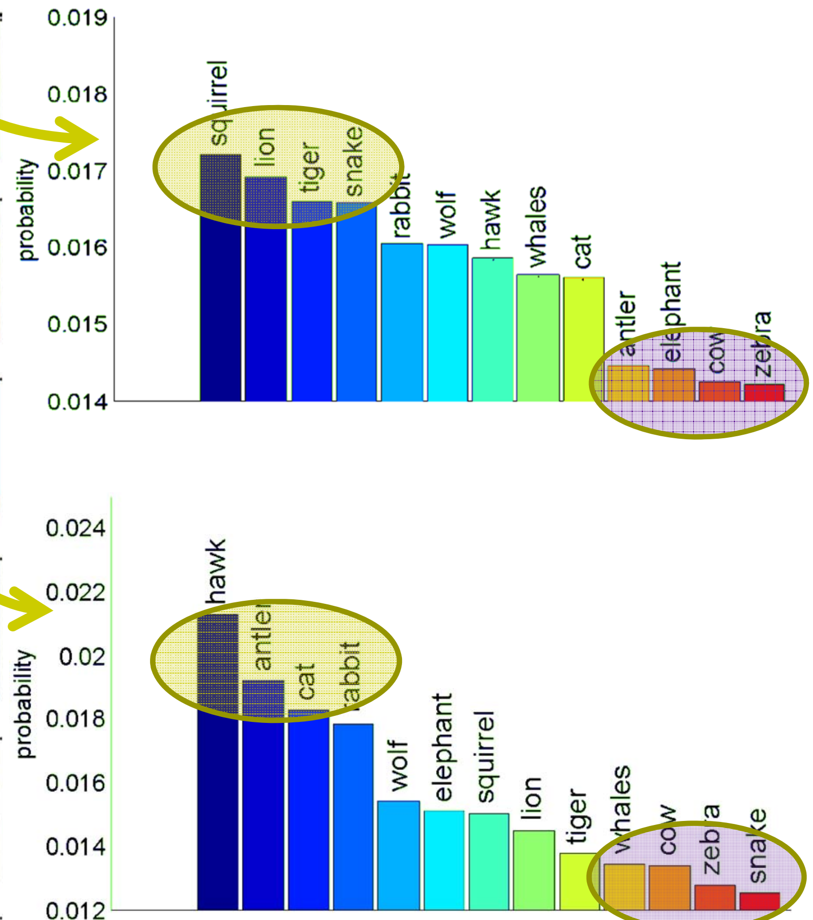


- Avg-KL: average pair-wise divergence

Predictive Latent Representation

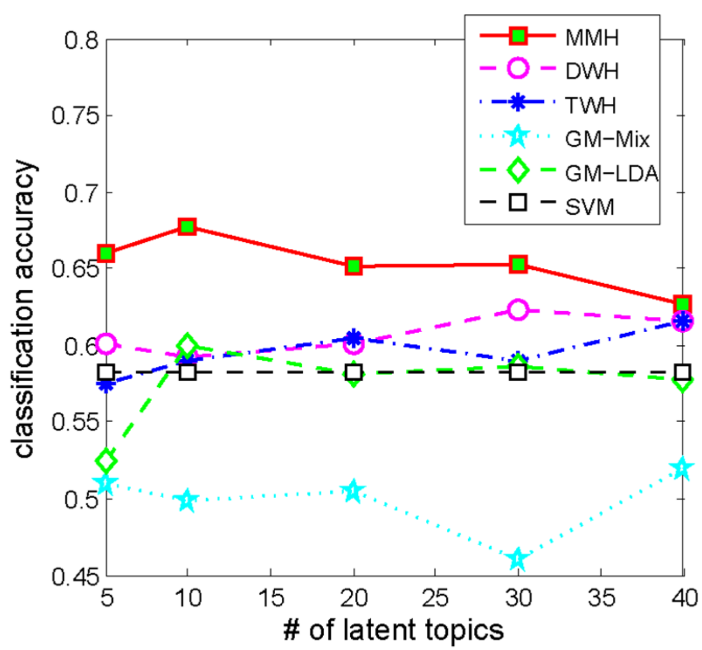
- Example latent topics discovered by a 60-topic MMH on Flickr Animal Data



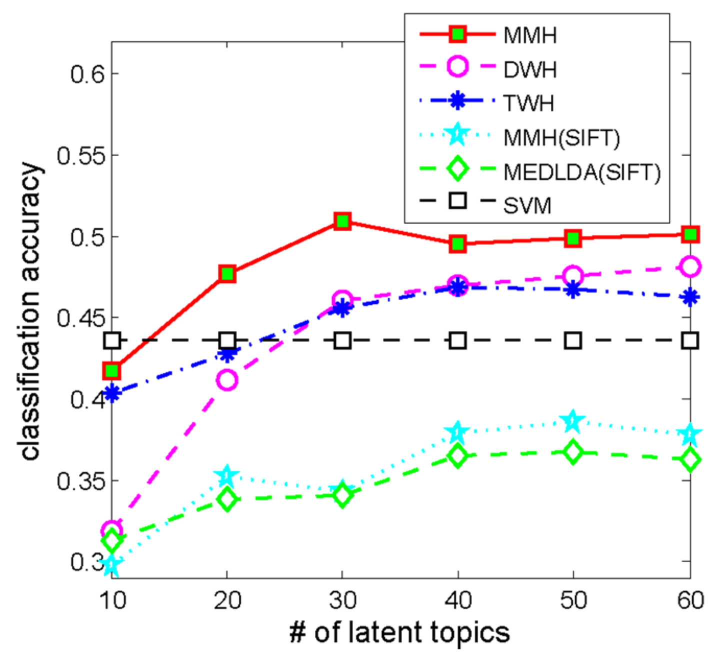


Classification Results

- **Data Sets:**
  - (Left) TRECVID 2003: (text + image features)
  - (Right) Flickr 13 Animal: (sift + image features)
- **Models**:
  - baseline(SVM),DWH+SVM, GM-Mixture+SVM, GM-LDA+SVM, TWH, MedLDA(sift only), MMH



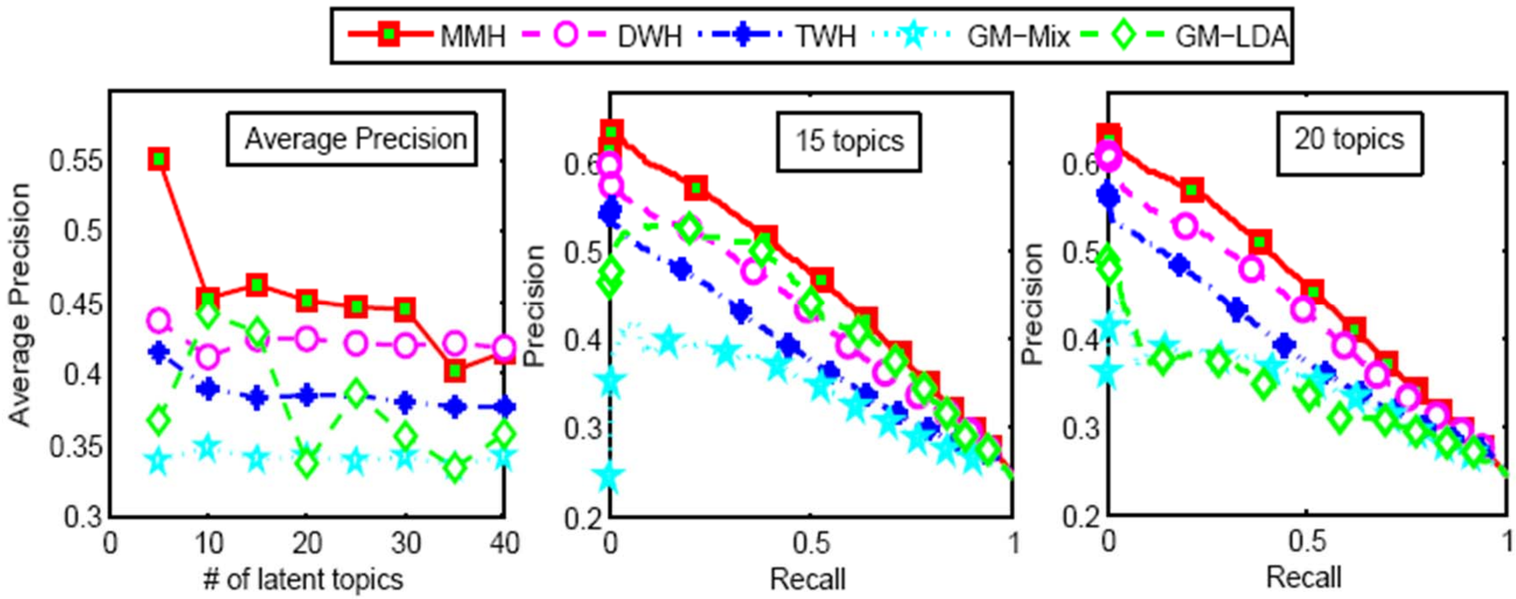
TRECVID



Flickr

Retrieval Results

- **Data Set: TRECVID 2003**
  - Each test sample is treated as a query, training samples are ranked based on the cosine similarity between a training sample and the given query
  - Similarity is computed based on the discovered latent topic representations
- **Models:** DWH, GM-Mixture, GM-LDA, TWH, MMH
- **Measure**: (Left) average precision on different topics and (Right) precision-recall curve



Infinite SVM and infinite latent SVM:

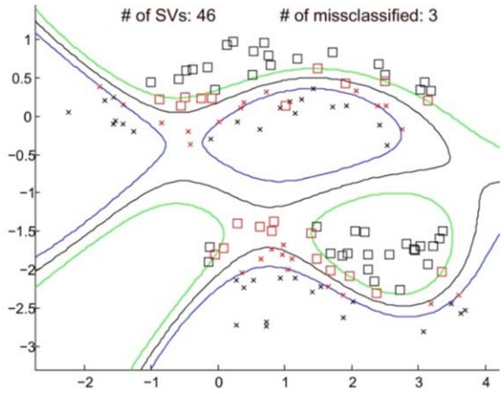
-- where SVMs meet NB for classification and feature selection

… where M is any combinations of classifiers and p is a nonparametric Bayesian prior

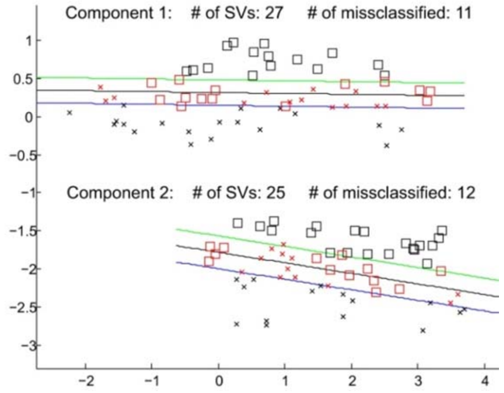
Mixture of SVMs

- Dirichlet process mixture of large-margin kernel machines
- Learn flexible non-linear local classifiers; potentially lead to a better control on model complexity, e.g., few unnecessary components

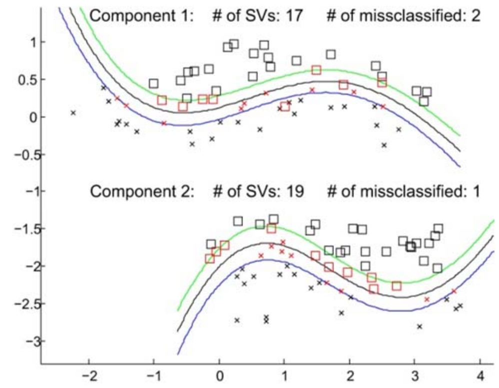
SVM using RBF kernel



Mixture of 2 linear SVM



Mixture of 2 RBF-SVM



- The first attempt to integrate Bayesian nonparametrics, large-margin learning, and kernel methods

Infinite SVM

- RegBayes framework:

$$\min_{p(\mathcal{M}),\xi} \quad \mathrm{KL}(p(\mathcal{M})\|\pi(\mathcal{M})) - \sum_{n=1}^{N}\int \log p(\mathbf{x}_n|\mathcal{M})p(\mathcal{M})d\mathcal{M} + U(\xi)$$

$$\text{s.t.}: \quad p(\mathcal{M}) \in \mathcal{P}_{\text{post}}(\xi),$$

convex function

direct and rich constraints on posterior distribution

- Model – latent class model
- Prior – Dirichlet process
- Likelihood – Gaussian likelihood
- Posterior constraints – max-margin constraints

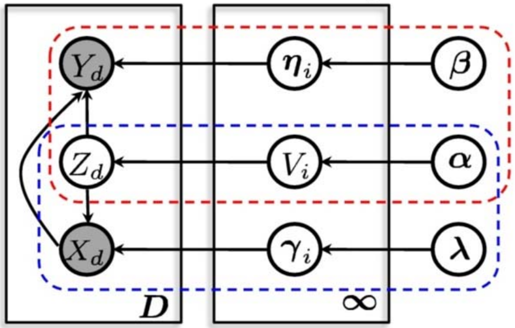
Infinite SVM

- DP mixture of large-margin classifiers

process of determining which classifier to use:

1. draw $V_i|\alpha \sim \text{Beta}(1, \alpha)$, $i \in \{1, 2, \cdots\}$.
2. draw $\eta_i|G_0 \sim G_0$, $i \in \{1, 2, \cdots\}$.
3. for the dth data point:

   (a) draw $Z_d|\{v_1, v_2, \cdots\} \sim \text{Mult}(\pi(\mathbf{v}))$



Graphical model with stick-breaking construction of DP

- Given a component classifier:

$$F(y, \mathbf{x}; z, \boldsymbol{\eta}) = \eta_z^\top \mathbf{f}(y, \mathbf{x}) = \sum_{i=1}^{\infty} \delta_{z,i} \eta_i^\top \mathbf{f}(y, \mathbf{x})$$

- Overall discriminant function:

$$F(y, \mathbf{x}) = \mathbb{E}_{q(z,\boldsymbol{\eta})}[F(y, \mathbf{x}; z, \boldsymbol{\eta})] == \sum_{i=1}^{\infty} q(z = i)\mathbb{E}_q[\eta_i]^\top \mathbf{f}(y, \mathbf{x})$$

- Prediction rule: $y^* = \arg\max_y F(y, \mathbf{x})$

- Learning problem: $\min_{q(\mathbf{z},\boldsymbol{\eta})} \text{KL}(q(\mathbf{z}, \boldsymbol{\eta}) \| p_0(\mathbf{z}, \boldsymbol{\eta})) + C_1 \mathcal{R}(q(\mathbf{z}, \boldsymbol{\eta}))$,

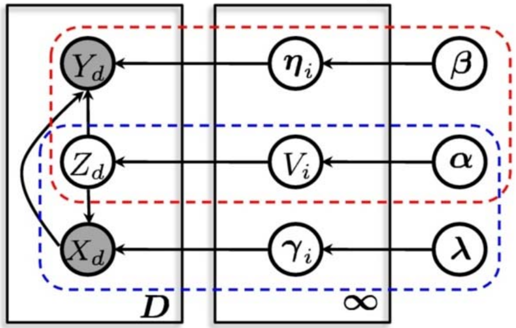
$$\mathcal{R}(q(\mathbf{z}, \boldsymbol{\eta})) = \sum_d \max_y (\ell_d^\Delta(y) + F(y, \mathbf{x}_d) - F(y_d, \mathbf{x}_d))$$

Infinite SVM

- Assumption and relaxation
  - Truncated variational distribution

$$q(\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \mathbf{v}) = \prod_{d=1}^{D} q(z_d) \prod_{t=1}^{T} q(\eta_t) \prod_{t=1}^{T} q(\gamma_t) \prod_{t=1}^{T-1} q(v_t)$$

  - Upper bound the KL-regularizer



Graphical model with stick-breaking construction of DP

- Opt. with coordinate descent
  - For $q(\boldsymbol{\eta})$, we solve an SVM learning problem
  - For $q(\mathbf{z})$, we get the closed update rule

$$q(z_d = t) \propto \exp\left\{(\mathbb{E}[\log v_t] + \sum_{i=1}^{t-1} \mathbb{E}[\log(1 - v_i)]) + \rho(\mathbb{E}[\gamma_t]^\top \mathbf{x}_d - \mathbb{E}[A(\gamma_t)]) + (1-\rho) \sum_y \omega_d^y \mu_t^\top \mathbf{f}_d^\Delta(y)\right\}$$

    - The last term regularizes the mixing proportions to favor prediction
  - For $q(\boldsymbol{\gamma}), q(\mathbf{v})$, the same update rules as in (Blei & Jordan, 2006)

Experiments on high-dim real data

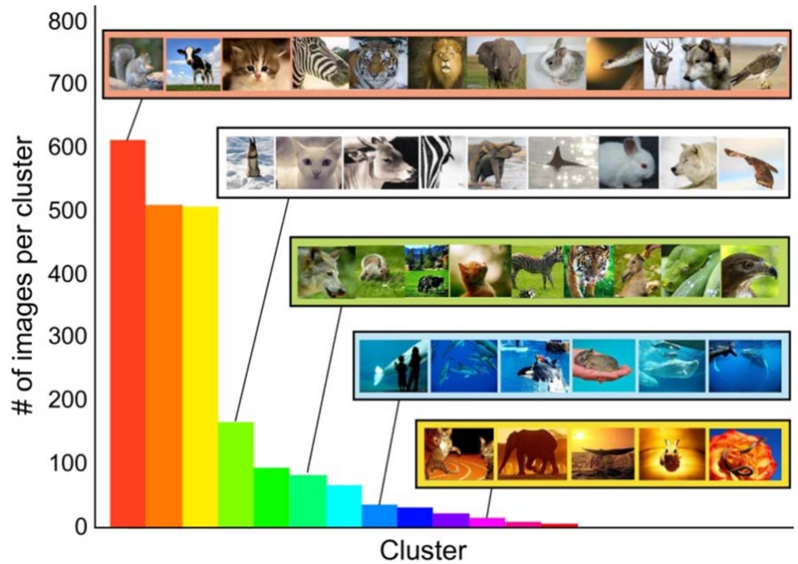
- Classification results and test time:

Table 4. Classification accuracy (%), F1 score (%), and test time (sec) for different models on the Flickr image dataset. All methods except dpMNL are implemented in C.

	ACCURACY	F1 SCORE	TEST TIME
MNL	49.8 ± 0.0	48.4 ± 0.0	**0.02** ± 0.00
MMH	51.7 ± 0.0	50.1 ± 0.0	0.33 ± 0.01
RBF-SVM	52.2 ± 0.0	48.4 ± 0.0	7.58 ± 0.06
DPMNL-EFH70	51.2 ± 0.9	49.9 ± 0.8	42.1 ± 7.39
DPMNL-PCA50	51.9 ± 0.7	49.9 ± 0.8	27.4 ± 2.08
LINEAR-ISVM	53.2 ± 0.4	51.3 ± 0.4	0.22 ± 0.01
RBF-ISVM	**54.2** ± 0.5	**51.6** ± 0.7	6.67 ± 0.05

For training, linear-iSVM is very efficient (~200s); RBF-iSVM is much slower, but can be significantly improved using efficient kernel methods (Rahimi & Recht, 2007; Fine & Scheinberg, 2001)

- Clusters:
  - simiar backgroud images group
  - a cluster has fewer categories



Learning Latent Features

- Infinite SVM is a Bayesian nonparametric latent class model
  - discover clustering structures
  - each data point is assigned to a single cluster/class

- Infinite Latent SVM is a Bayesian nonparametric latent feature/factor model
  - discover latent factors
  - each data point is mapped to a set (can be infinite) of latent factors

  - Latent factor analysis is a key technique in many fields; Popular models are FA, PCA, ICA, NMF, LSI, etc.

Infinite Latent SVM

- RegBayes framework:

$$\min_{p(\mathcal{M}),\xi} \quad \mathrm{KL}(p(\mathcal{M})\|\pi(\mathcal{M})) - \sum_{n=1}^{N} \int \log p(\mathbf{x}_n|\mathcal{M})p(\mathcal{M})d\mathcal{M} + U(\xi)$$

$$\mathrm{s.t.:} \quad p(\mathcal{M}) \in \mathcal{P}_{\mathrm{post}}(\xi),$$

convex function

direct and rich constraints on posterior distribution

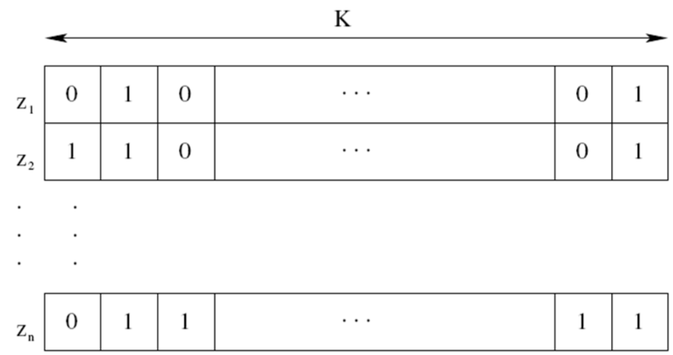
- Model – latent feature model
- Prior – Indian Buffet process
- Likelihood – Gaussian likelihood
- Posterior constraints – max-margin constraints

Beta-Bernoulli Latent Feature Model

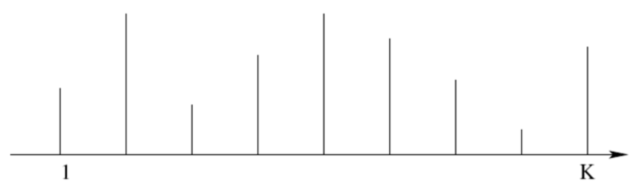
- A random finite binary latent feature models

$$\pi_k | \alpha \sim \text{Beta}(\frac{\alpha}{K}, 1)$$

$$z_{ik} | \pi_k \sim \text{Bernoulli}(\pi_k)$$



  - π_k is the relative probability of each feature being on, e.g.,

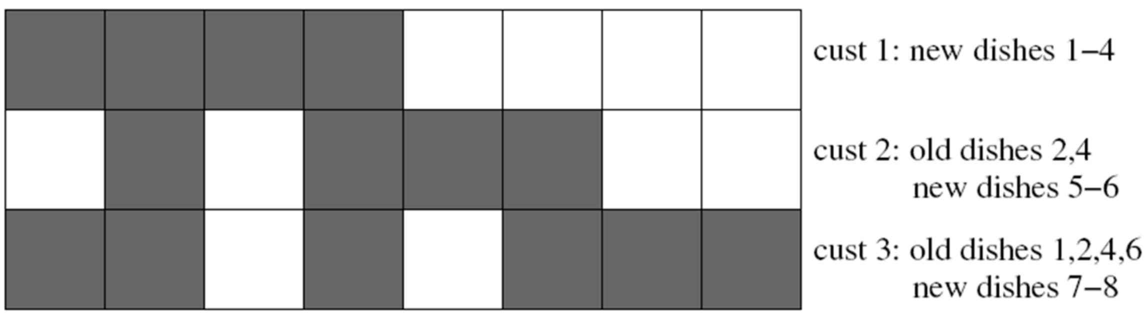


  - $z_{i\cdot}$ are binary vectors, giving the latent structure that's used to generate the data, e.g.,

$$\mathbf{x}_i \sim \mathcal{N}(\eta^\top z_{i\cdot}, \delta^2)$$

Indian Buffet Process

- A stochastic process on infinite binary feature matrices
- Generative procedure:
  - Customer 1 chooses the first K_1 dishes: $K_1 \sim \mathrm{Poisson}(\alpha)$
  - Customer *i* chooses:
    - Each of the existing dishes with probability $\frac{m_k}{i}$
    - K_i additional dishes, where $K_i \sim \mathrm{Poisson}(\frac{\alpha}{i})$



$$Z_{i\cdot} \sim \mathcal{IBP}(\alpha)$$

Posterior Constraints – classification

- Suppose latent features **z** are given, we define *latent discriminant function*:

$$f(y, \mathbf{x}, \mathbf{z}; \boldsymbol{\eta}) = \boldsymbol{\eta}^\top \mathbf{g}(y, \mathbf{x}, \mathbf{z})$$

- Define *effective discriminant function* (reduce uncertainty):

$$f(y, \mathbf{x}; p(\mathbf{Z}, \boldsymbol{\eta})) = \mathbb{E}_{p(\mathbf{Z}, \boldsymbol{\eta})}[f(y, \mathbf{x}, \mathbf{z}; \boldsymbol{\eta})] = \mathbb{E}_{p(\mathbf{Z}, \boldsymbol{\eta})}[\boldsymbol{\eta}^\top \mathbf{g}(y, \mathbf{x}, \mathbf{z})]$$

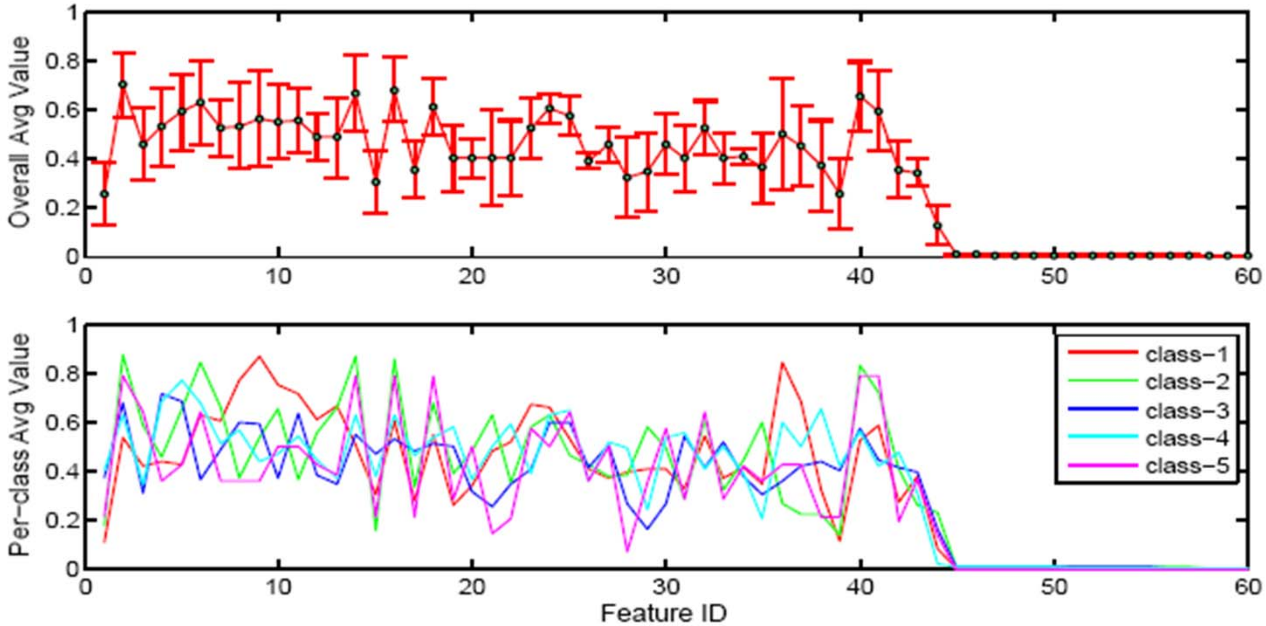
- Posterior constraints with max-margin principle

$$\forall n \in \mathcal{I}_{\mathrm{tr}}, \forall y: \quad f(y_n, \mathbf{x}_n; p(\mathbf{Z}, \boldsymbol{\eta})) - f(y, \mathbf{x}_n; p(\mathbf{Z}, \boldsymbol{\eta})) \geq \ell(y, y_n) - \xi_n$$

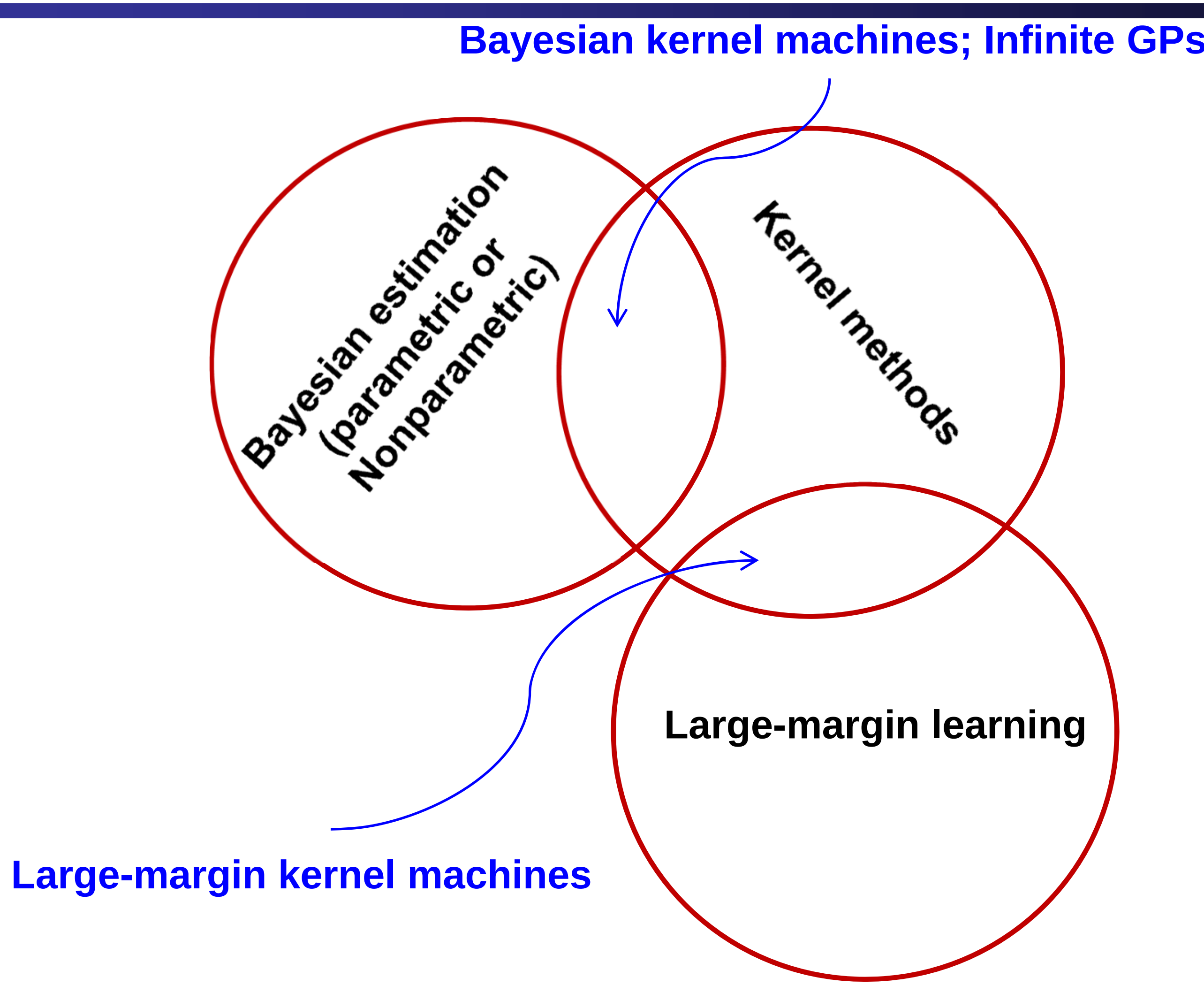
Experimental Results

- Classification
  - Accuracy and F1 scores on TRECVID2003 and Flickr image datasets

Model	TRECVID2003 Accuracy	TRECVID2003 F1 score	Flickr Accuracy	Flickr F1 score
EFH+SVM	0.565 ± 0.0	0.427 ± 0.0	0.476 ± 0.0	0.461 ± 0.0
MMH	$\mathbf{0.566} \pm 0.0$	0.430 ± 0.0	$\mathbf{0.538} \pm 0.0$	$\mathbf{0.512} \pm 0.0$
IBP+SVM	0.553 ± 0.013	0.397 ± 0.030	0.500 ± 0.004	0.477 ± 0.009
iLSVM	0.563 ± 0.010	$\mathbf{0.448} \pm 0.011$	0.533 ± 0.005	0.510 ± 0.010

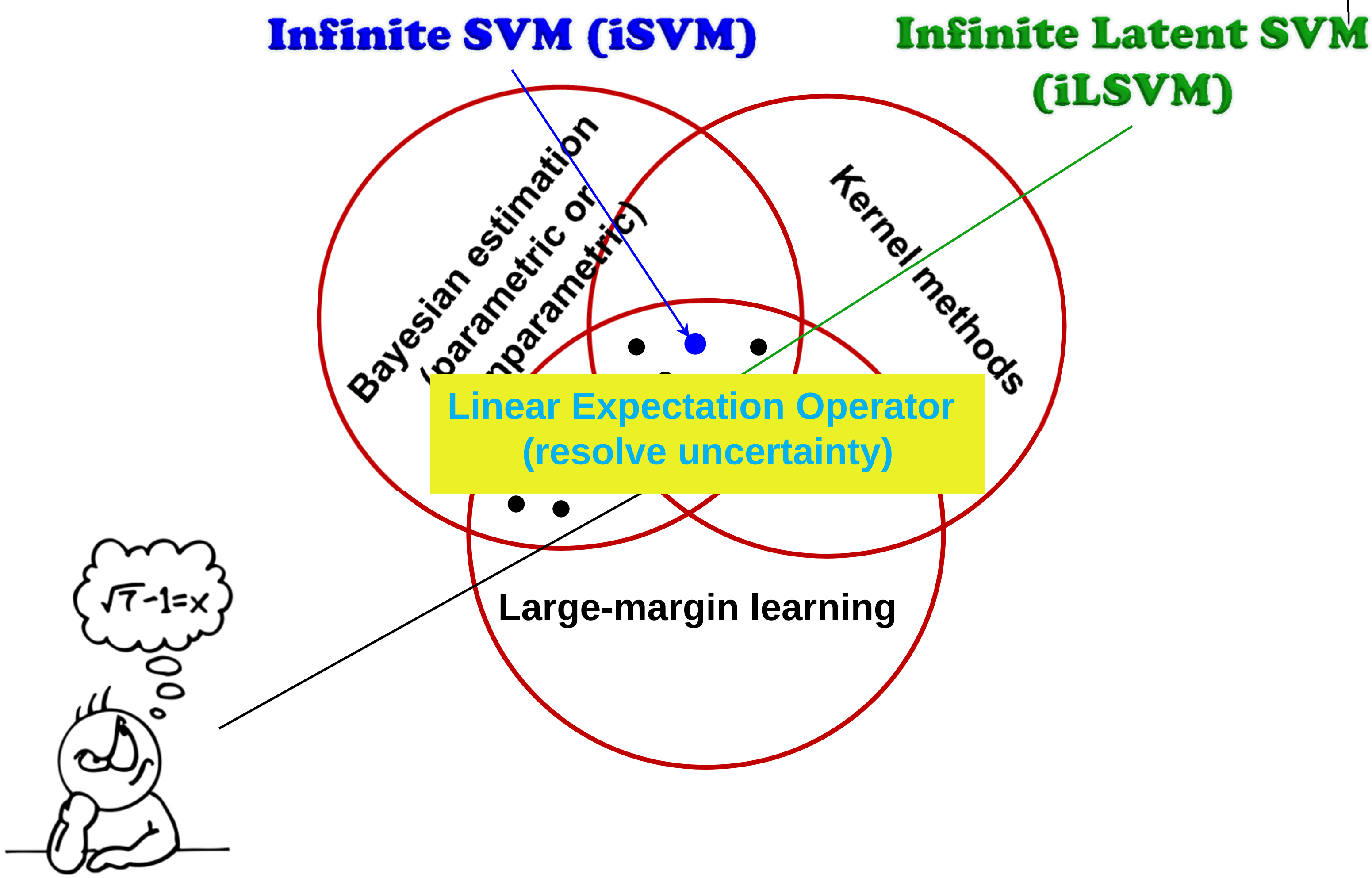


Summary



Bayesian kernel machines; Infinite GPs
Bayesian estimation (parametric or Nonparametric)
Kernel methods
Large-margin learning
Large-margin kernel machines

Summary



Infinite SVM (iSVM)
Infinite Latent SVM (iLSVM)
Bayesian estimation
parametric or
Kernel methods
Linear Expectation Operator (resolve uncertainty)
Large-margin learning
√7-1=x

Summary

- A general framework of MaxEnDNet for learning structured input/output models
  - Subsumes the standard M^3Ns
  - Model averaging: PAC-Bayes theoretical error bound
  - Entropic regularization: sparse M^3Ns
  - **Generative + discriminative: latent variables, semi-supervised learning on partially labeled data, fast inference**
- PoMEN
  - Provides an elegant approach to incorporate latent variables and structures under max-margin framework
  - Enable Learning arbitrary graphical models discriminatively
- Predictive Latent Subspace Learning
  - MedLDA for text topic learning
  - Med total scene model for image understanding
  - Med latent MNs for multi-view inference
- Bayesian nonparametrics meets max-margin learning
- Experimental results show the advantages of max-margin learning over likelihood methods in **EVERY** case.

Remember: Elements of Learning

- **Here are some important elements to consider before you start:**
  - Task:
    - Embedding? Classification? Clustering? Topic extraction? …
  - Data and other info:
    - Input and output (e.g., continuous, binary, counts, …)
    - Supervised or unsupervised, of a blend of everything?
    - Prior knowledge? Bias?
  - Models and paradigms:
    - BN? MRF? Regression? SVM?
    - Bayesian/Frequents ? Parametric/Nonparametric?
  - Objective/Loss function:
    - MLE? MCLE? Max margin?
    - Log loss, hinge loss, square loss? …
  - Tractability and exactness trade off:
    - Exact inference? MCMC? Variational? Gradient? Greedy search?
    - Online? Batch? Distributed?
  - Evaluation:
    - Visualization? Human interpretability? Perperlexity? Predictive accuracy?
- **It is better to consider one element at a time!**