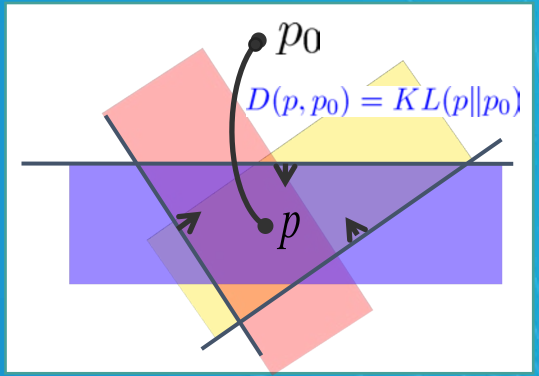
Probabilistic Graphical Models

Integrative Paradigms of GMs: Regularized Bayesian Methods

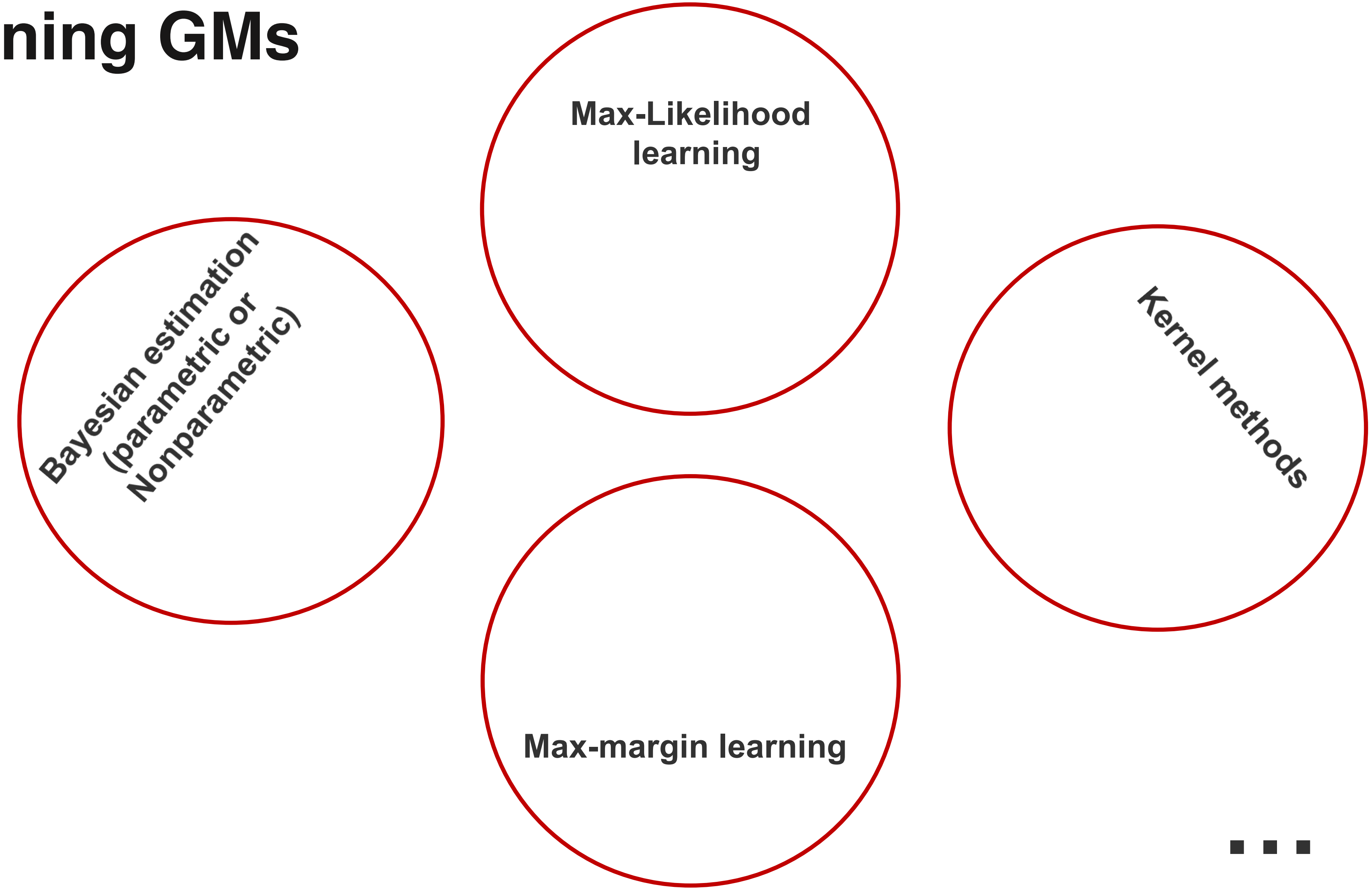
Eric Xing

Lecture 24, April 15, 2019

Reading: see class homepage

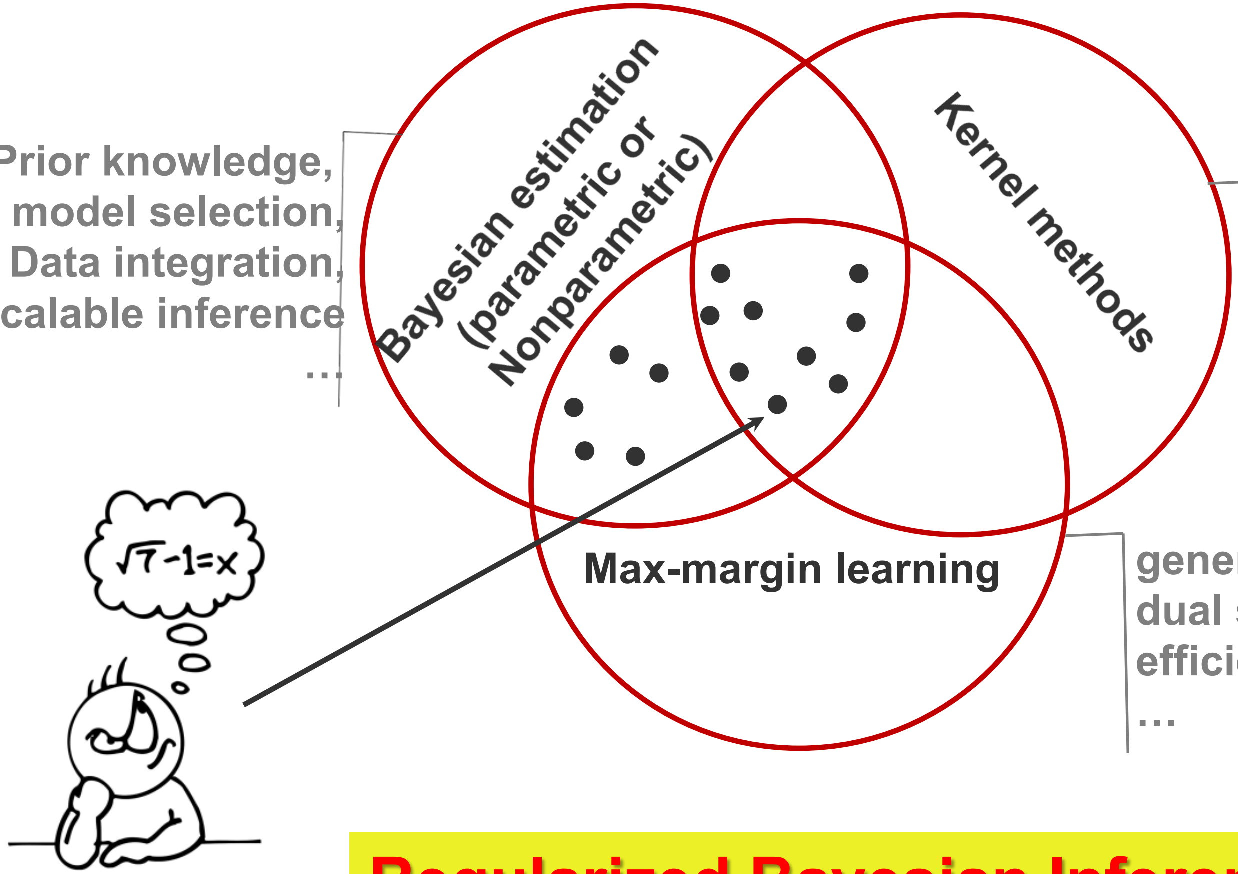


Learning GMs



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

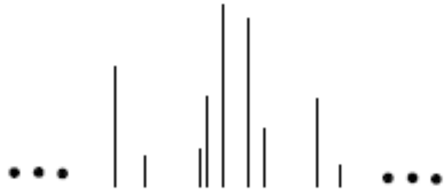
$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}} \propto p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})$$

Examples:

→ see next slide

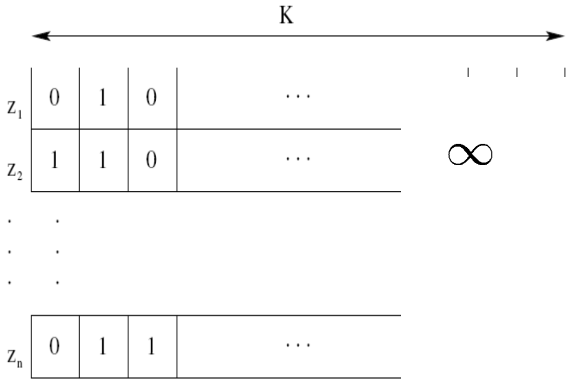
Nonparametric Bayesian Inference

probability measure



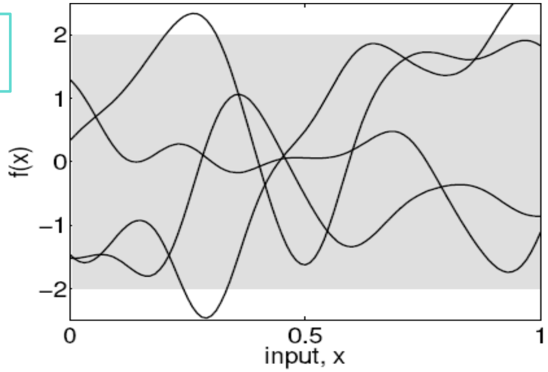
Dirichlet Process Prior [Antoniak, 1974]
+ Multinomial/Gaussian/Softmax likelihood

binary matrix



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

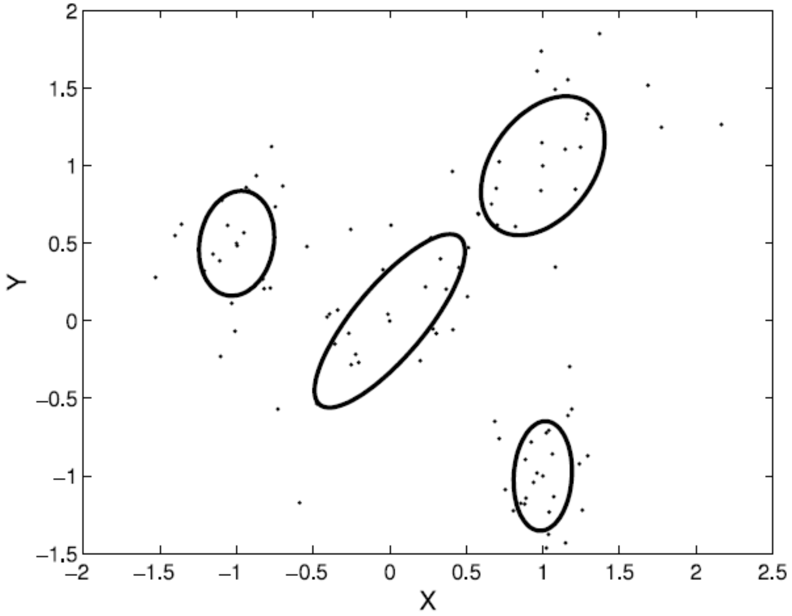
function



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



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:

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

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.,

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

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?

MLE versus max-margin learning

- Likelihood-based estimation
  - Probabilistic (joint/conditional likelihood model)
  - Easy to perform Bayesian learning, and incorporate prior knowledge, latent structures, missing data
  - Bayesian or direct regularization
  - Hidden structures or generative hierarchy

- Max-margin learning
  - Non-probabilistic (concentrate on input-output mapping)
  - Not obvious how to perform Bayesian learning or consider prior, and missing data
  - Support vector property, sound theoretical guarantee with limited samples
  - Kernel tricks

- Maximum Entropy Discrimination (MED) (Jaakkola, et al., 1999)
  - Model averaging $\hat{y} = \mathrm{sign} \int p(\mathbf{w}) F(x; \mathbf{w})\, d\mathbf{w}$ $\quad (y \in \{+1, -1\})$
  - The optimization problem (binary classification)

$$\min_{p(\Theta)} \; KL(p(\Theta) || p_0(\Theta))$$

$$\text{s.t.} \; \int p(\Theta)[y_i F(x; \mathbf{w}) - \xi_i]\, d\Theta \geq 0, \forall i,$$

where Θ is the parameter $\mathbf{w}$ when ξ are kept fixed or the pair $(\mathbf{w}, \xi)$ when we want to optimize over ξ

Classical Predictive Models

- Input and output space: $\mathcal{X} \triangleq \mathbb{R}^{M_x}$ $\mathcal{Y} \triangleq \{-1, +1\}$
- Learning: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathcal{W}} \ell(\mathrm{x}, y; \mathbf{w}) + \lambda R(\mathbf{w})$

where $\ell(\cdot)$ represents a convex loss, and $R(\mathbf{w})$ is a regularizer preventing overfitting

– Logistic Regression
- Max-likelihood (or MAP) estimation

$$\max_{\mathbf{w}} \mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \sum_{i=1}^{N} \log p(y^i | \mathrm{x}^i; \mathbf{w}) + \mathcal{N}(\mathbf{w})$$

- Corresponds to a **Log loss** with L2 R

$$\ell_{LL}(\mathrm{x}, y; \mathbf{w}) \triangleq \ln \sum_{y' \in \mathcal{Y}} \exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y')\} - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y)$$

– Support Vector Machines (SVM)
- Max-margin learning

$$\min_{\mathbf{w}, \xi} \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C \sum_{i=1}^{N} \xi_i;$$

$$\text{s.t. } \forall i, \forall y' \neq y^i : \mathbf{w}^\top \Delta \mathbf{f}_i(y') \geq 1 - \xi_i, \ \xi_i \geq 0.$$

- Corresponds to a **hinge loss** with L2 R

$$\ell_{MM}(\mathrm{x}, y; \mathbf{w}) \triangleq \max_{y' \in \mathcal{Y}} \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y') - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y) + \ell'(y', y)$$

Advantages:
1. Full probabilistic semantics
2. Straightforward Bayesian or direct regularization
3. Hidden structures or generative hierarchy

Advantages:
1. Dual sparsity: few support vectors
2. Kernel tricks
3. Strong empirical results

Structured Prediction Graphical Models

- Input and output space: $\mathcal{X} \triangleq \mathbb{R}_{X_1} \times, \ldots, \mathbb{R}_{X_K} \quad \mathcal{Y} \triangleq \mathbb{R}_{Y_1} \times, \ldots, \mathbb{R}_{Y_{K'}}$

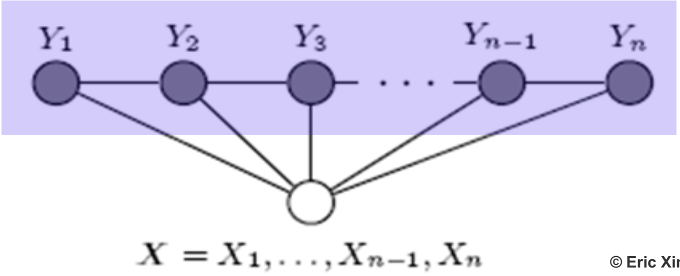
- Conditional Random Fields (CRFs) (Lafferty et al 2001)
  - Based on a Logistic Loss (LR)
  - Max-likelihood estimation (point-estimate)

$$\mathcal{L}(\mathcal{D};\mathbf{w}) \triangleq \log \sum_{\mathbf{y}'} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}')) - \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}) + R(\mathbf{w})$$

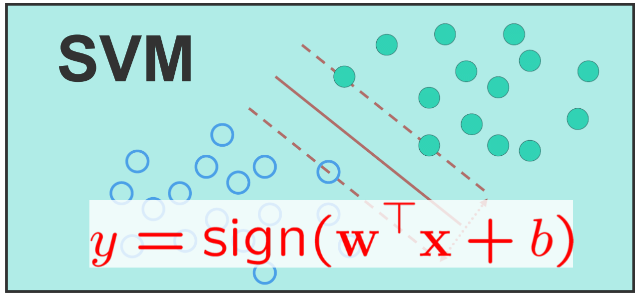
- **Max-margin Markov Networks (M^3Ns) (Taskar et al 2003)**
  - **Based on a Hinge Loss (SVM)**
  - **Max-margin learning (point-estimate)**

$$\mathcal{L}(\mathcal{D};\mathbf{w}) \triangleq \log \max_{\mathbf{y}'} \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}') - \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}) + \ell(\mathbf{y}',\mathbf{y}) + R(\mathbf{w})$$

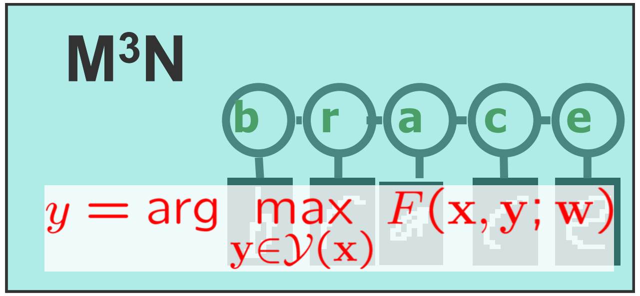
- Markov properties are encoded in the feature functions $\mathbf{f}(\mathbf{x},\mathbf{y})$



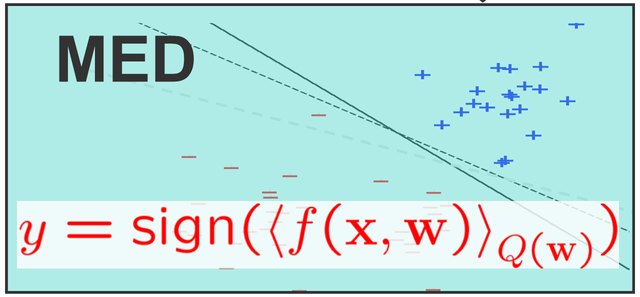
Max-Margin Learning Paradigms



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



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



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

MED-MN ?

= SMED + “Bayesian” M^{3}N

Maximum Entropy Discrimination Markov Networks

(Zhu et al, ICML 2008)

- Structured MaxEnt Discrimination (SMED):

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

generalized maximum entropy or *regularized* KL-divergence

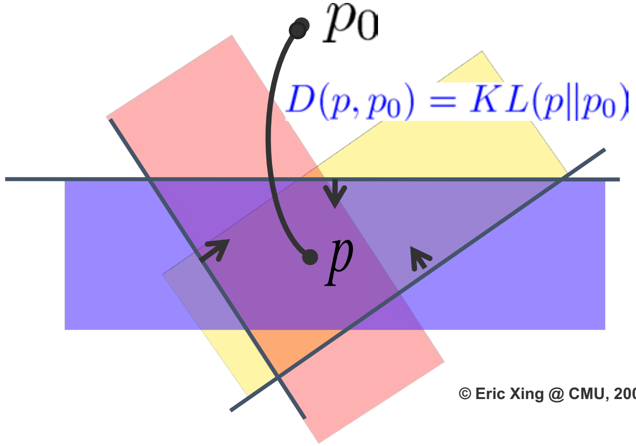
- Feasible subspace of weight distribution:

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

expected margin constraints.

- Average from distribution of M^{3}Ns

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



Solution to MaxEnDNet

- Theorem:
  - Posterior Distribution:

$$p(\mathbf{w}) = \frac{1}{Z(\alpha)} p_0(\mathbf{w}) \exp\{\sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})[\Delta F_i(\mathbf{y};\mathbf{w}) - \Delta \ell_i(\mathbf{y})]\}$$

  - Dual Optimization Problem:

$$\begin{aligned} \mathrm{D1}: \quad & \max_{\alpha} \; -\log Z(\alpha) - U^\star(\alpha) \\ & \text{s.t.} \;\; \alpha_i(\mathbf{y}) \geq 0, \; \forall i, \; \forall \mathbf{y}, \end{aligned}$$

$U^\star(\cdot)$ is the conjugate of the $U(\cdot)$, i.e., $U^\star(\alpha) = \sup_\xi \left(\sum_{i,y} \alpha_i(y)\xi_i - U(\xi) \right)$

Algorithmic issues of solving M³Ns

- **Primal problem:**

$$\text{P0 } (\text{M}^3\text{N}): \min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t. } \forall i, \forall \mathbf{y} \neq \mathbf{y}^i: \mathbf{w}^\top \Delta \mathbf{f}_i(\mathbf{y}) \geq \Delta \ell_i(\mathbf{y}) - \xi_i,$$
$$\xi_i \geq 0,$$

- **Algorithms**
  - Cutting plane
  - Sub-gradient
  - …

- **Dual problem:**

$$\text{D0 } (\text{M}^3\text{N}): \max_{\alpha} \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y}) - \frac{1}{2}\eta^\top\eta$$

$$\text{s.t. } \forall i, \ \forall \mathbf{y}: \sum_{\mathbf{y}} \alpha_i(\mathbf{y}) = C; \ \alpha_i(\mathbf{y}) \geq 0.$$

$$\text{where } \eta = \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y}).$$

- **Algorithms:**
  - SMO
  - Exponentiated gradient
  - …

- **Nonlinear Features with Kernels**
  - Generative entropic kernels [Martins et al, JMLR 2009]
  - Nonparametric RKHS embedding of rich distributions [on going]
- **Approximate decoders for global features**
  - LP-relaxed Inference (polyhedral outer approx.) [Martins et al, ICML 09, ACL 09]
  - Balancing Accuracy and Runtime: Loss-augmented inference

Variational Learning of LapMEDN

- Exact primal or dual function is hard to optimize

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

$$\max_{\alpha}\ \sum_{i,\mathbf{y}}\alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y})-\sum_{k=1}^{K}\log\frac{\lambda}{\lambda-\eta_k^2}$$
$$\text{s.t.}\ \ \sum_{\mathbf{y}}\alpha_i(\mathbf{y})=C;\ \alpha_i(\mathbf{y})\geq 0,\ \forall i,\ \forall\mathbf{y}.$$

- Use the hierarchical representation of Laplace prior, we get:

$$KL(p||p_0) = -H(p) - \langle\log\int p(\mathbf{w}|\tau)p(\tau|\lambda)\,\mathrm{d}\tau\,\rangle_p$$
$$\leq -H(p) - \langle\int q(\tau)\log\frac{p(\mathbf{w}|\tau)p(\tau|\lambda)}{q(\tau)}\,\mathrm{d}\tau\,\rangle_p \triangleq \mathcal{L}(p(\mathbf{w}),q(\tau))$$

- We optimize an upper bound:

$$\min_{p(\mathbf{w})\in\mathcal{F}_1;q(\tau);\xi}\ \mathcal{L}(p(\mathbf{w}),q(\tau))+U(\xi)$$

- Why is it easier?
  - Alternating minimization leads to nicer optimization problems

Keep $q(\tau)$ fixed	Keep $p(\mathbf{w})$ fixed
- The effective prior is normal $\forall k:\ p_0(w_k\|\tau_k)=\mathcal{N}(w_k\|0,\langle\frac{1}{\tau_k}\rangle_{q(\tau)}^{-1})$	- Closed form solution of $q(\tau)$ and its expectation $\langle\frac{1}{\tau_k}\rangle_q=\sqrt{\frac{\lambda}{\langle w_k^2\rangle_p}}.$

An M[3]N optimization problem!

Closed-form solution!

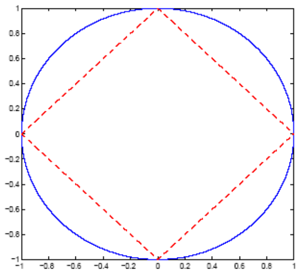
The 3 advantages of MEDN

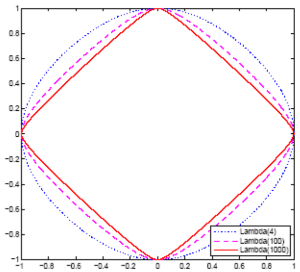
- An averaging Model: PAC-Bayesian prediction error guarantee (Theorem 3)

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

- Entropy regularization: Introducing useful biases
  - Standard Normal prior => reduction to standard M^3N (we've seen it)
  - Laplace prior => Posterior shrinkage effects (sparse M^3N)

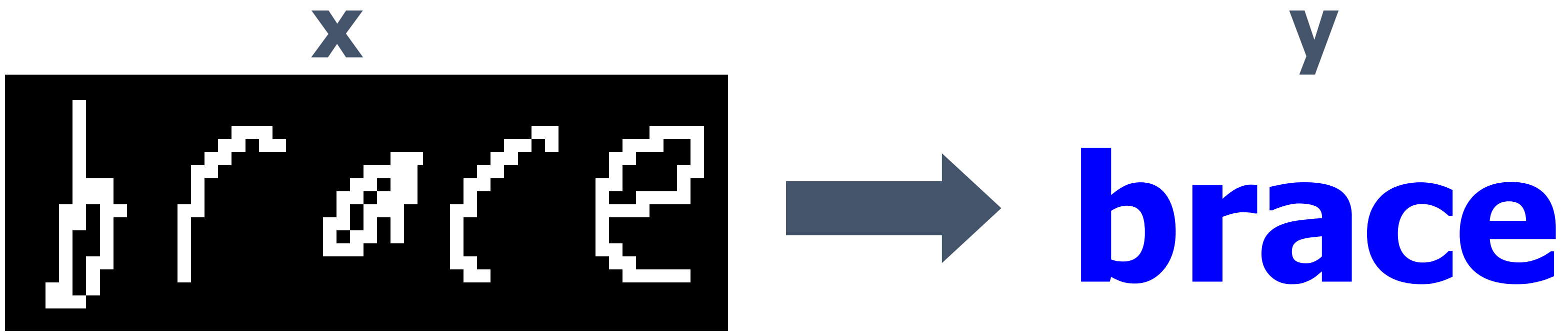
$$\min_{\mu,\xi} \sqrt{\lambda}\sum_{k=1}^{K}\Big(\sqrt{\mu_k^2 + \frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}}\log\frac{\sqrt{\lambda\mu_k^2+1}+1}{2}\Big) + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t. } \mu^\top \Delta\mathbf{f}_i(\mathbf{y}) \ge \Delta\ell_i(\mathbf{y}) - \xi_i;\ \xi_i \ge 0,\ \ \forall i,\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



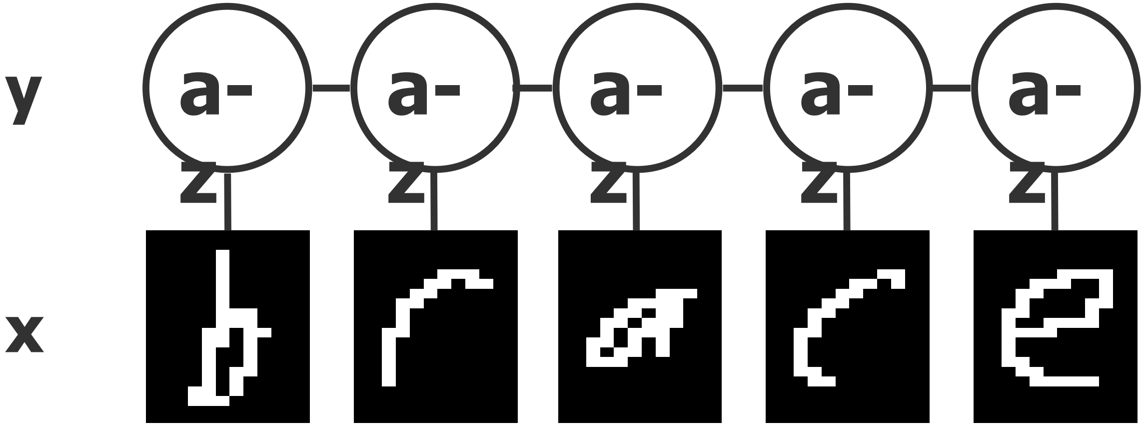


- Integrating Generative and Discriminative principles
- Incorporate latent variables and structures (PoMEN)
  - Semisupervised learning (with partially labeled data)

Experimental results on OCR datasets



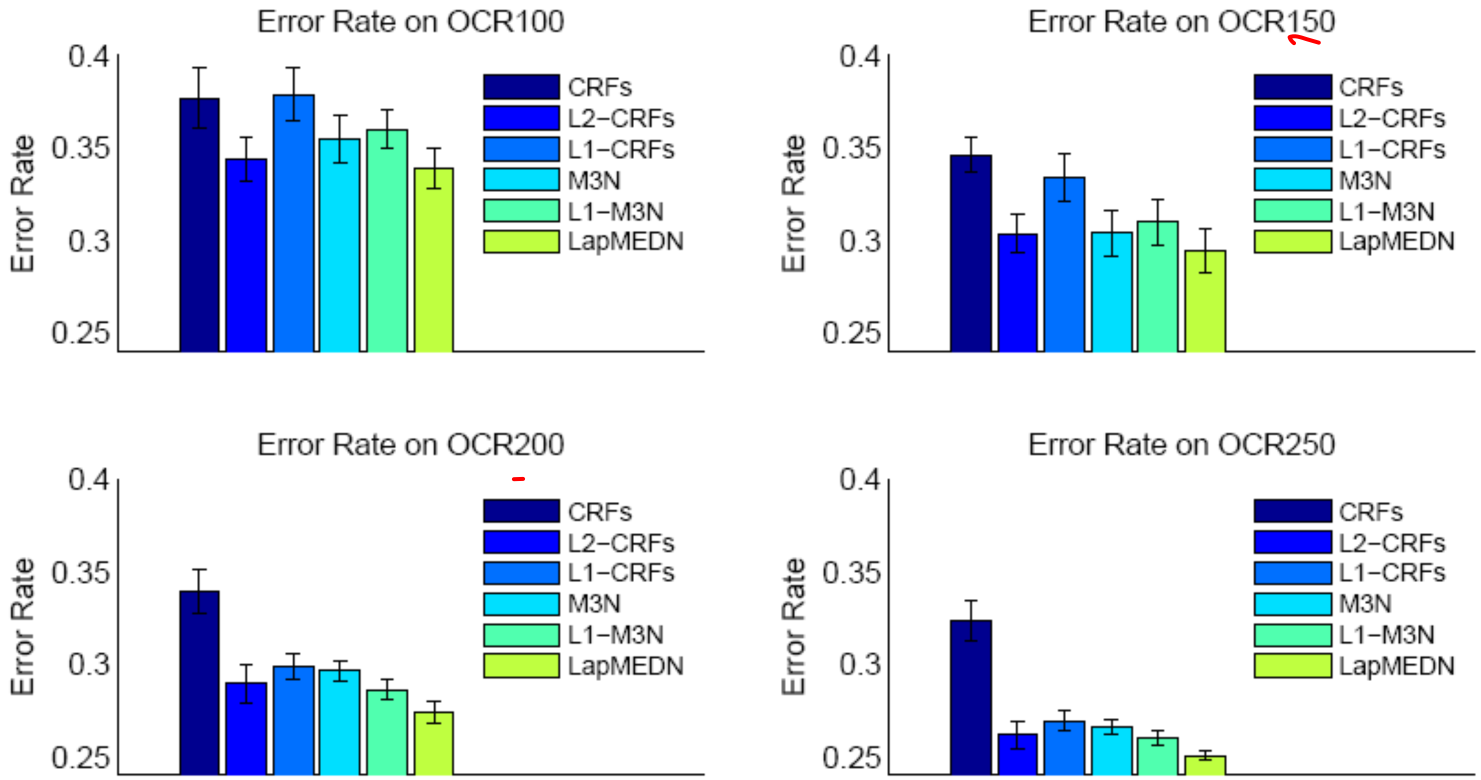
Structured output



Experimental results on OCR datasets

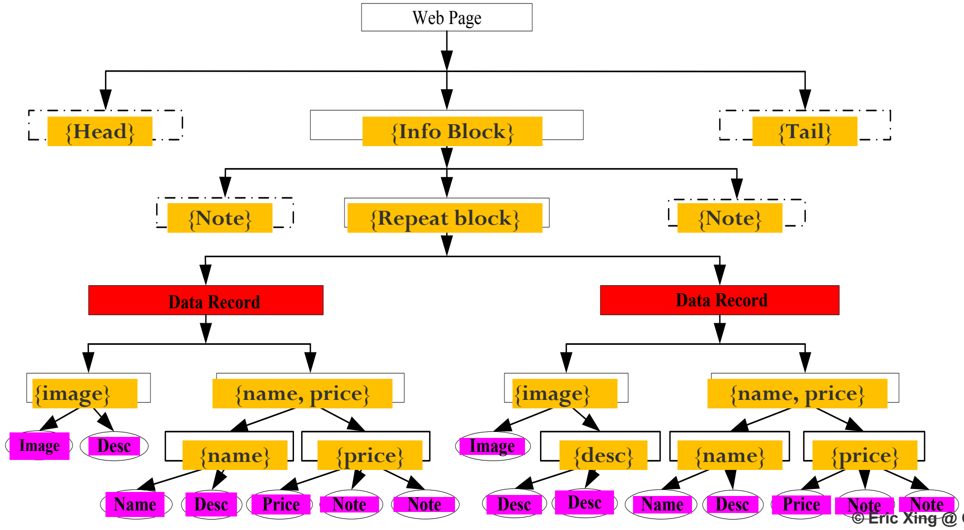
$$\left(\text{CRFs},\ L_1-\text{CRFs},\ L_2-\text{CRFs},\ \text{M}^3\text{Ns},\ L_1-\text{M}^3\text{Ns},\ \text{and LapMEDN}\right)$$

- We randomly construct OCR100, OCR150, OCR200, and OCR250 for 10 fold CV.



Latent Hierarchical MaxEnDNet

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

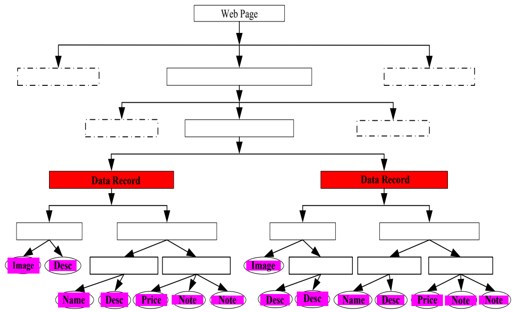


- **Hierarchical labeling**
- **Advantages:**
  - Computational efficiency
  - Long-range dependency
  - Joint extraction

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 $p(\mathbf{w}, \mathbf{z})$



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

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

  - Prediction:

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

Alternating Minimization Alg.

- Factorization assumption:

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

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

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

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

  - **Normal prior**
    - **M³N problem (QP)**
  - **Laplace prior**
    - **Laplace M³N problem (VB)**

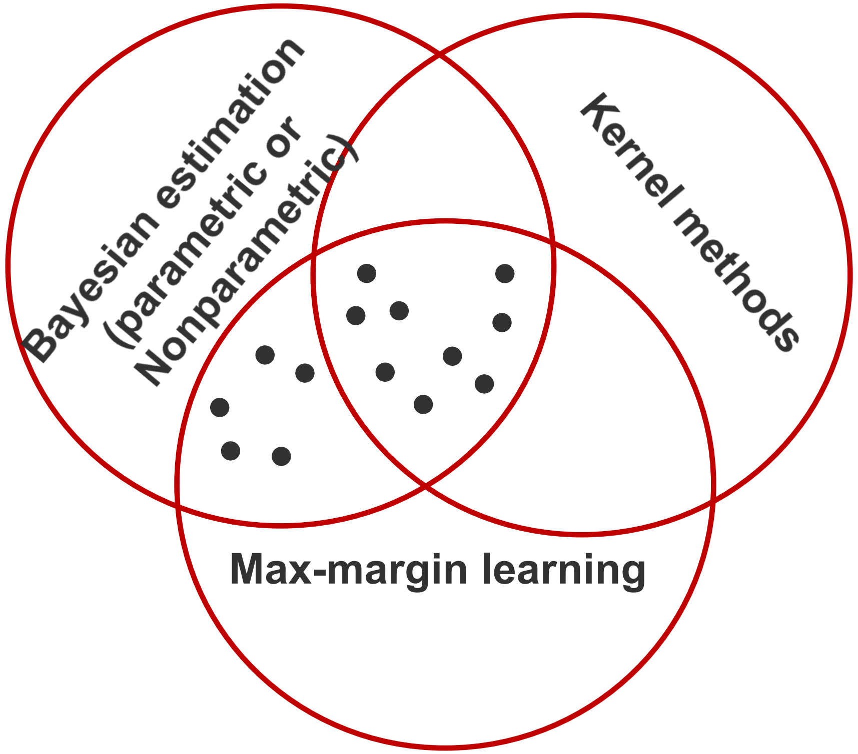
  - Step 2: keep $p(\mathbf{w})$ fixed, optimize over $p(\mathbf{z})$

$$\min_{p(\mathbf{w}),\xi} \; KL(p(\mathbf{z})||p_0(\mathbf{z})) + C\xi_i$$
$$\text{s.t.} \;\; p(\mathbf{z}) \in \mathcal{F}^\star_1, \;\; \xi_i \geq 0.$$

$$\mathcal{F}^\star_1 = \{p(\mathbf{z}) : \sum_{\mathbf{z}} p(\mathbf{z}) \int p(\mathbf{w})[\Delta F_i(\mathbf{y}, \mathbf{z}; \mathbf{w}) - \Delta \ell_i(\mathbf{y})] \, \mathrm{d}\mathbf{w} \geq -\xi_i, \; \forall i, \; \forall \mathbf{y}\}$$

Equivalently reduced to an LP with a polynomial number of constraints

An integrative paradigm for learning GM --- RegBayes



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

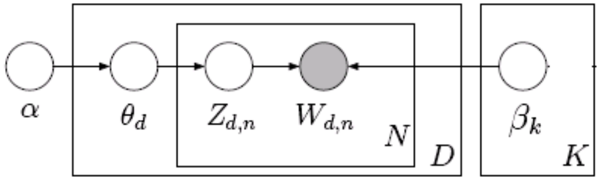
Predictive Latent Subspace Learning

via a large-margin approach

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

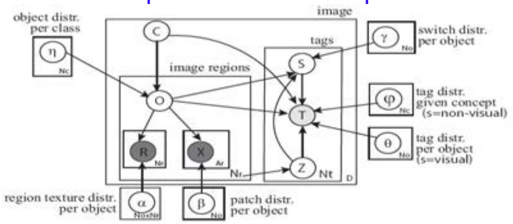
Unsupervised Latent Subspace Discovery

- 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]



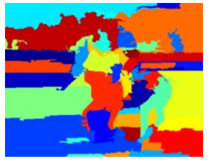
⇒

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

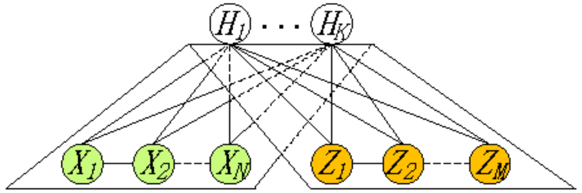


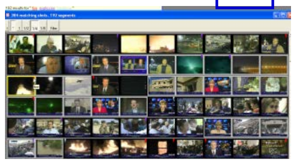
Athlete
Horse
Grass
Trees
Sky
Saddle

⇒

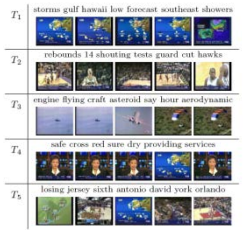


  - Multi-view latent Markov models [Xing et al 2005]





⇒



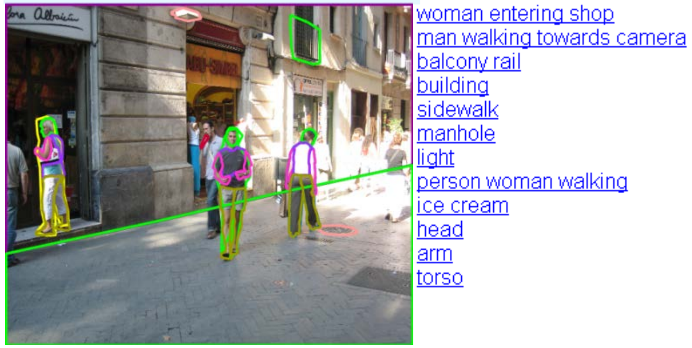
  - PCA, CCA, …

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)

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



- Many others

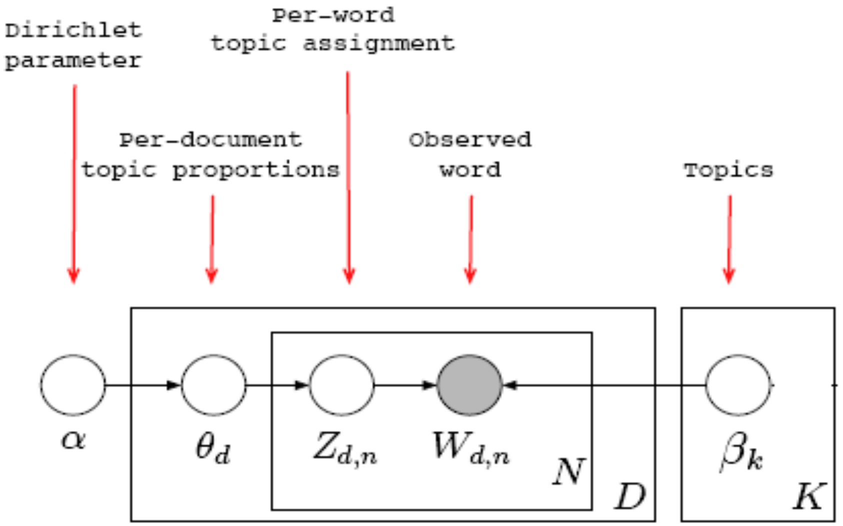
Flickr (http://www.flickr.com/)

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: $$p(\theta, \mathbf{z}, \mathbf{W}|\alpha, \beta) = \prod_{d=1}^{D} p(\theta_d|\alpha)\left(\prod_{n=1}^{N} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)\right)$$ **exact inference intractable!**

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

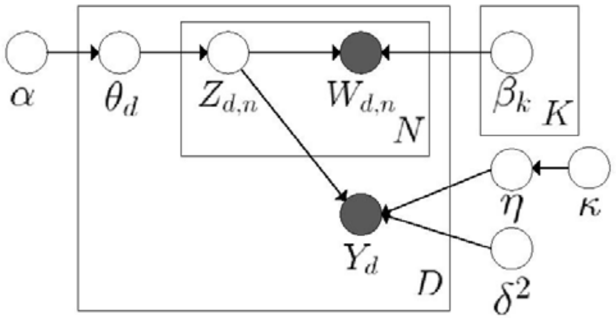
$$\mathcal{L}(q) \triangleq -E_q[\log p(\theta, \mathbf{z}, \mathbf{W}|\alpha, \beta)] - \mathcal{H}(q(\mathbf{z}, \theta)) \geq -\log p(\mathbf{W}|\alpha, \beta)$$

- 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

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

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

model fitting

predictive accuracy

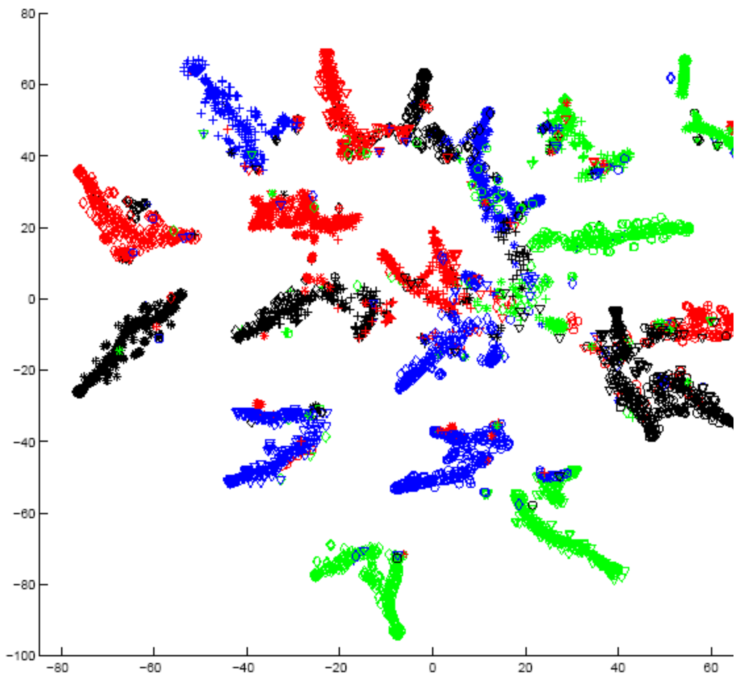
  - MedLDA Classification Model

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

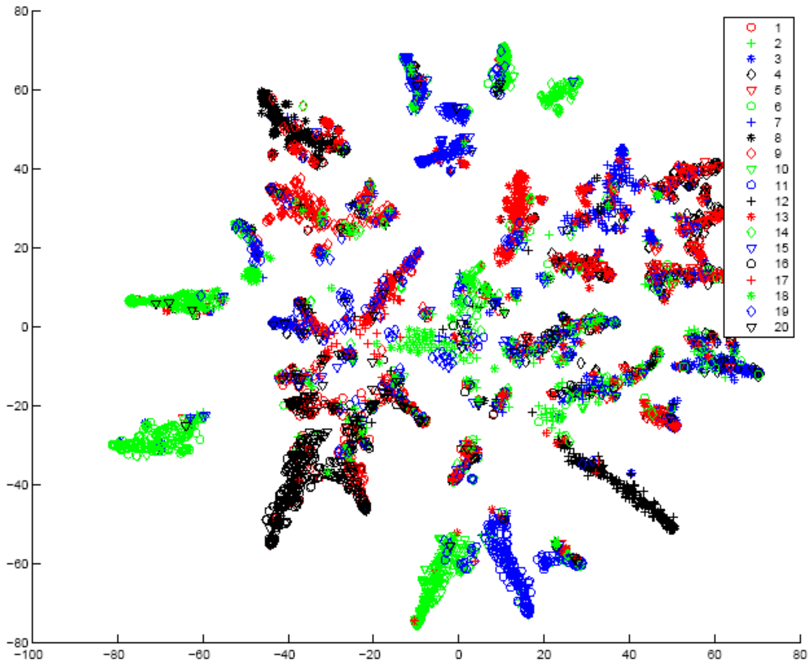
$$\text{s.t. } \forall d, \ y \neq y_d: \quad E[\eta^\top \Delta \mathbf{f}_d(y)] \geq 1 - \xi_d; \ \xi_d \geq 0.$$

Document Modeling

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



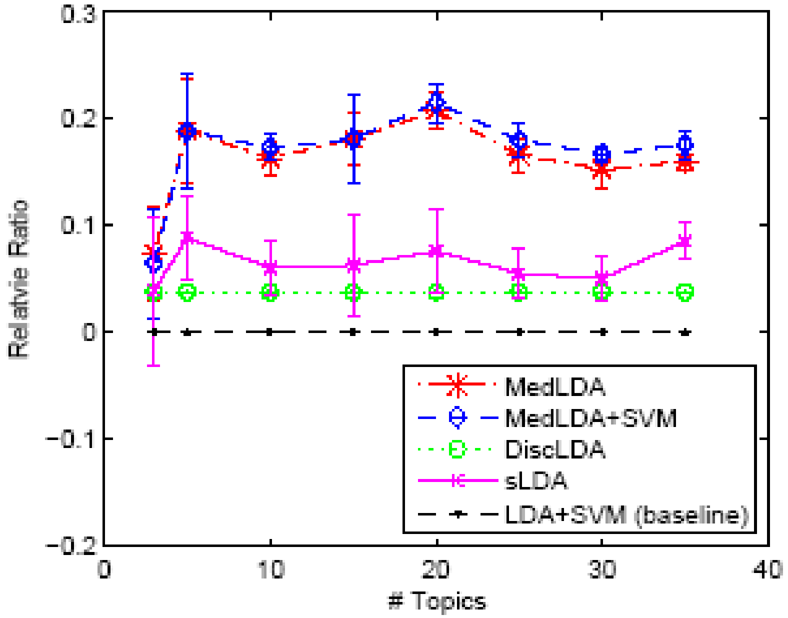
MedLDA

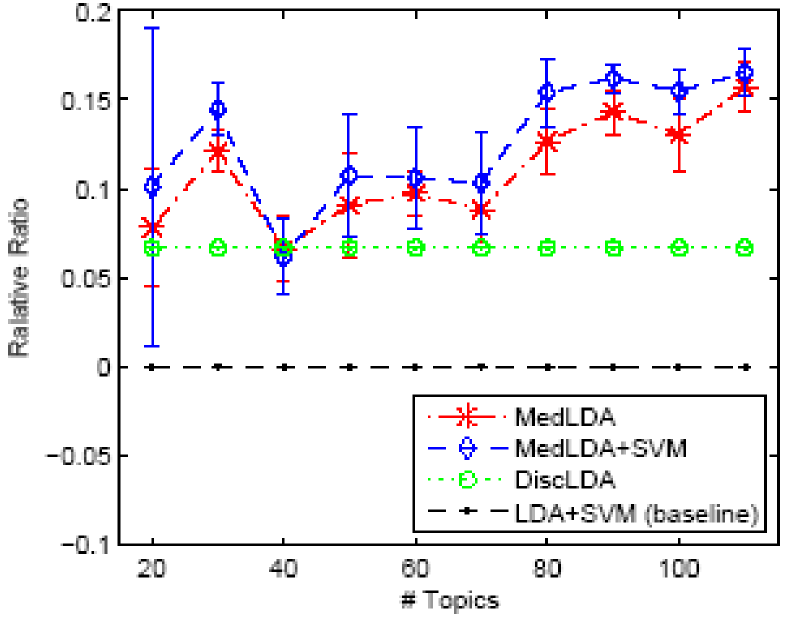


LDA

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



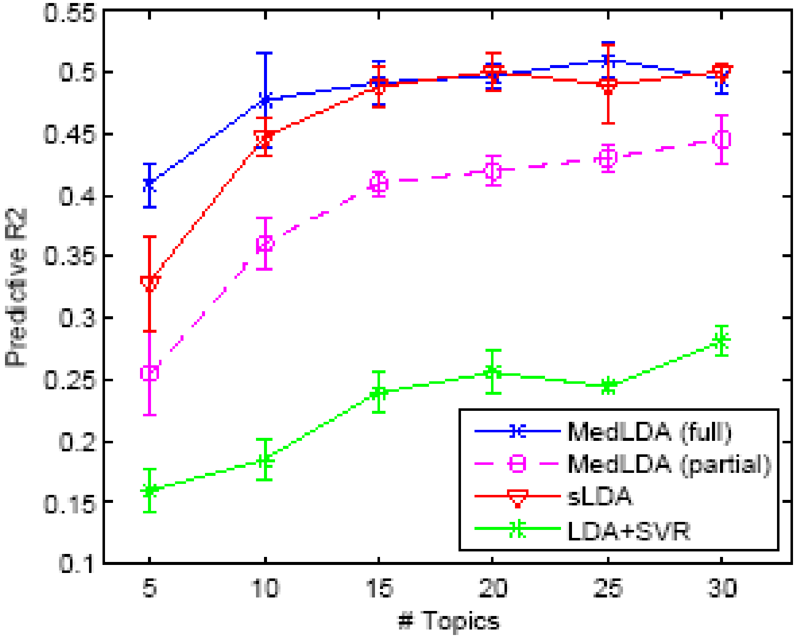


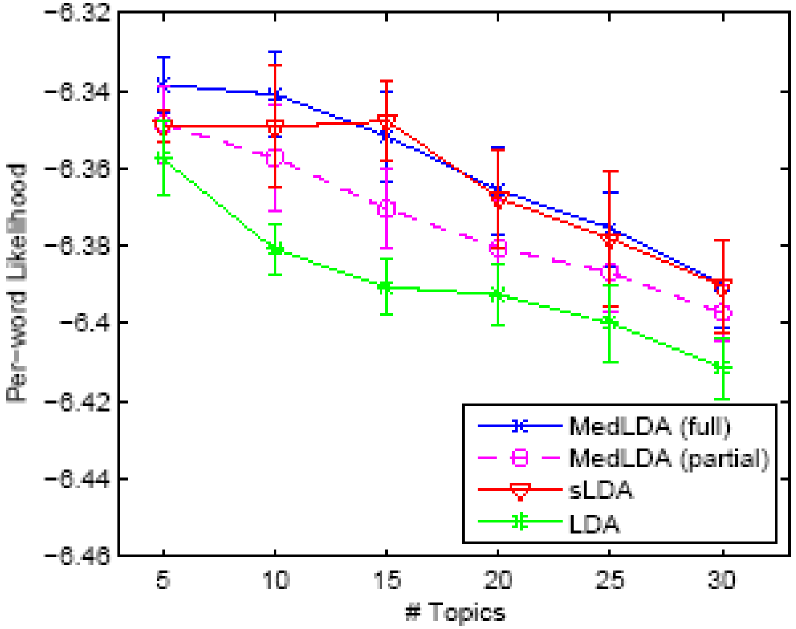
Regression

- **Data Set**: Movie Review (Blei & McAuliffe, 2007)
- **Models**: MedLDA(*partial*), MedLDA(*full*), sLDA, LDA+SVR

- **Measure**: predictive R^2 and per-word log-likelihood

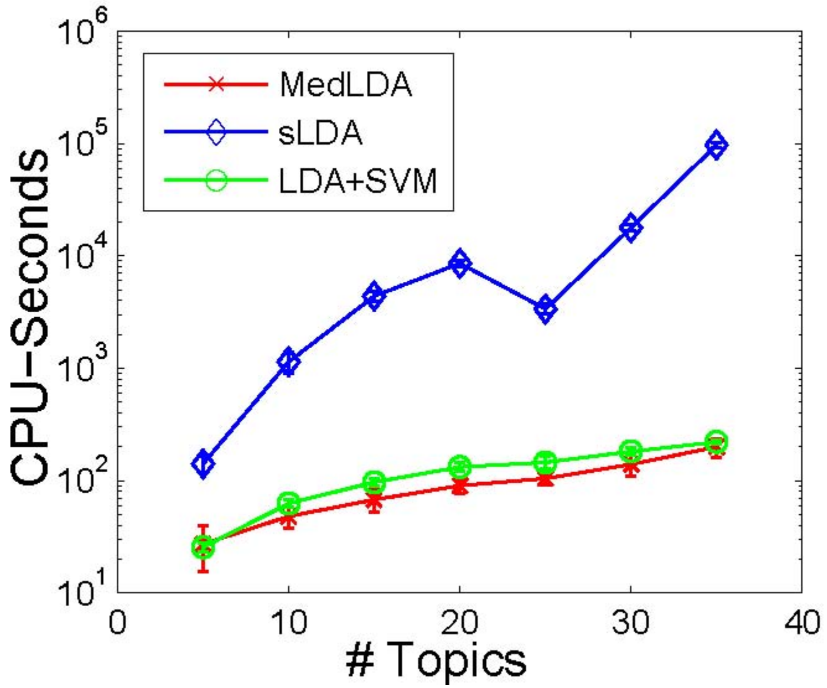
$$pR^2 = 1 - \frac{\sum_d (y_d - \hat{y}_d)^2}{\sum_d (y_d - \bar{y}_d)^2}$$





Time Efficiency

- Binary Classification



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

Infinite SVM and infinite latent SVM:

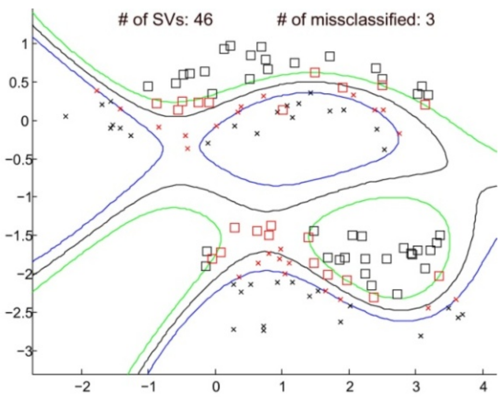
-- where SVMs meet NB for classification and feature selection

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

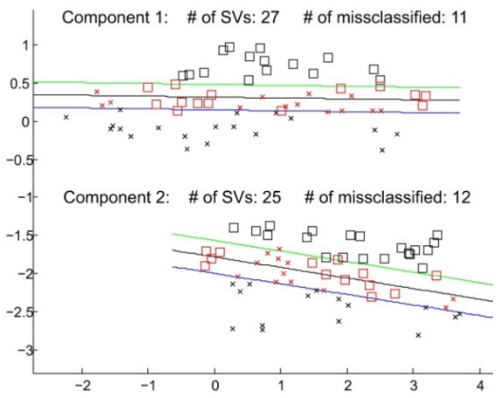
Mixture of SVMs

- 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

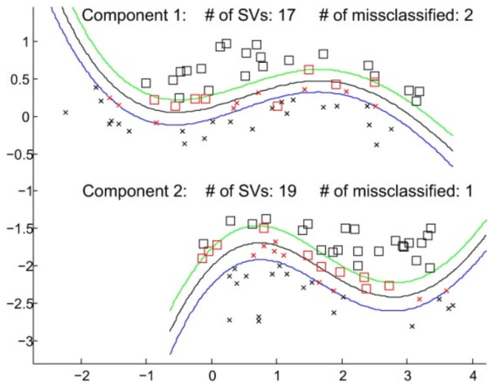
SVM using RBF kernel



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:

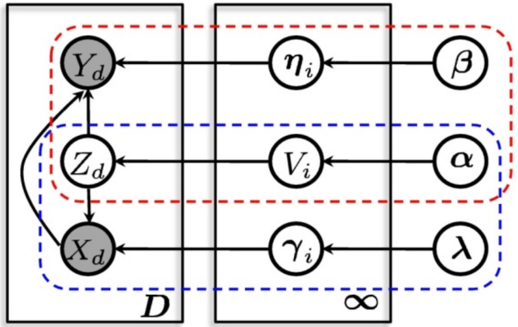
$$\min_{p(\mathcal{M}),\xi} \quad \mathrm{KL}(p(\mathcal{M})\|\pi(\mathcal{M})) - \sum_{n=1}^{N}\int \log p(\mathbf{x}_n|\mathcal{M})p(\mathcal{M})d\mathcal{M} + U(\xi)$$
$$\mathrm{s.t.}: \quad p(\mathcal{M}) \in \mathcal{P}_{\mathrm{post}}(\xi),$$

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



Graphical model with stick-breaking construction of DP

- DP mixture of large-margin classifiers

  process of determining which classifier to use:

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

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

  - 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:

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

  - Prediction rule:

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

  - Learning problem:

$$\min_{q(\mathbf{z},\boldsymbol{\eta})} \mathrm{KL}(q(\mathbf{z}, \boldsymbol{\eta}) \| p_0(\mathbf{z}, \boldsymbol{\eta})) + C_1 \mathcal{R}(q(\mathbf{z}, \boldsymbol{\eta})),$$

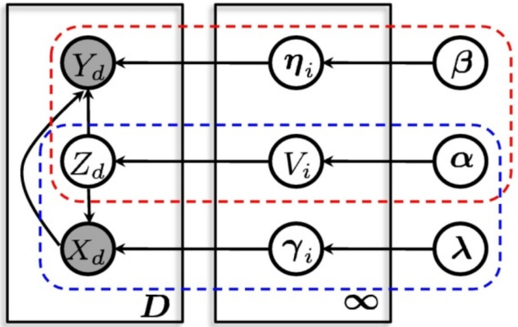
$$\mathcal{R}(q(\mathbf{z}, \boldsymbol{\eta})) = \sum_d \max_y (\ell_d^\Delta(y) + F(y, \mathbf{x}_d) - F(y_d, \mathbf{x}_d))$$

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

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

    - 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

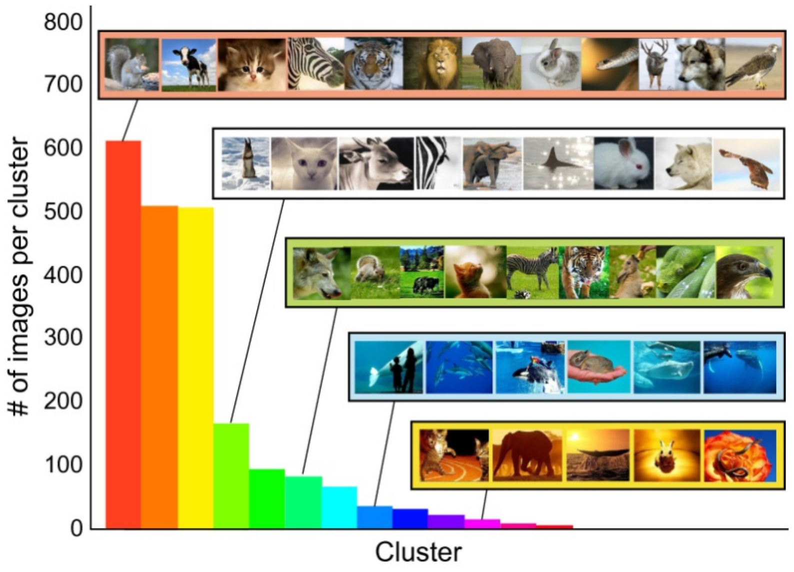
- Classification results and test time:

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	$\mathbf{0.02} \pm 0.00$
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	$\mathbf{54.2} \pm 0.5$	$\mathbf{51.6} \pm 0.7$	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:

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

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

convex function

direct and rich constraints on posterior distribution

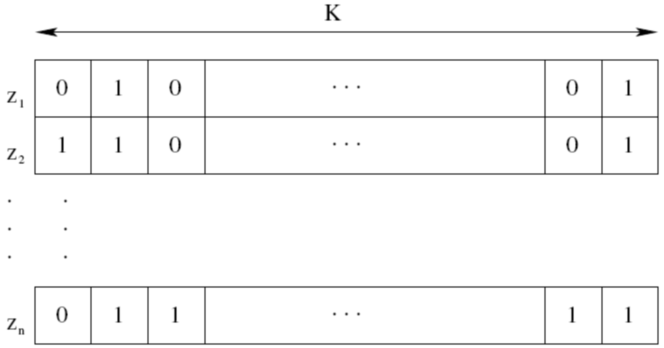
  - Model – latent feature model
  - Prior – Indian Buffet process
  - Likelihood – Gaussian likelihood
  - Posterior constraints – max-margin constraints

Beta-Bernoulli Latent Feature Model

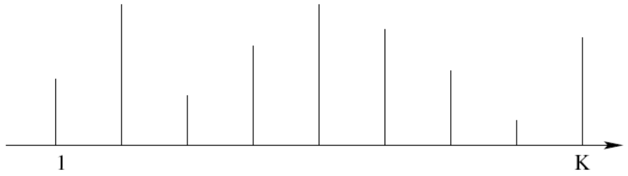
- A random finite binary latent feature models

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

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



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

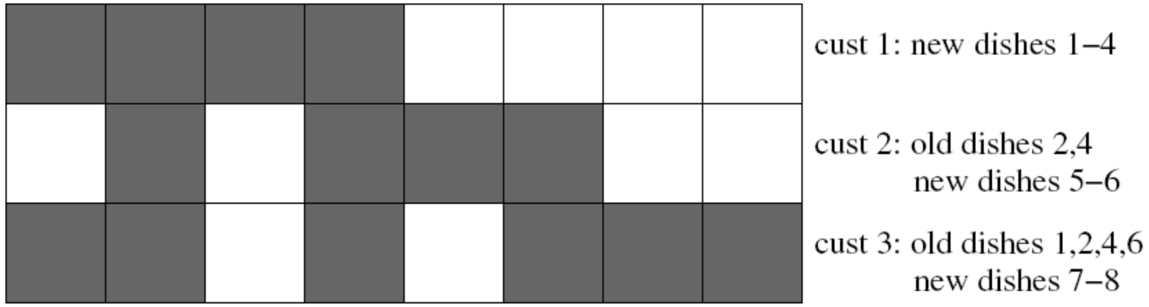


  - $z_{i\cdot}$ are binary vectors, giving the latent structure that's used to generate the data, e.g.,

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

Indian Buffet Process

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



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

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):

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

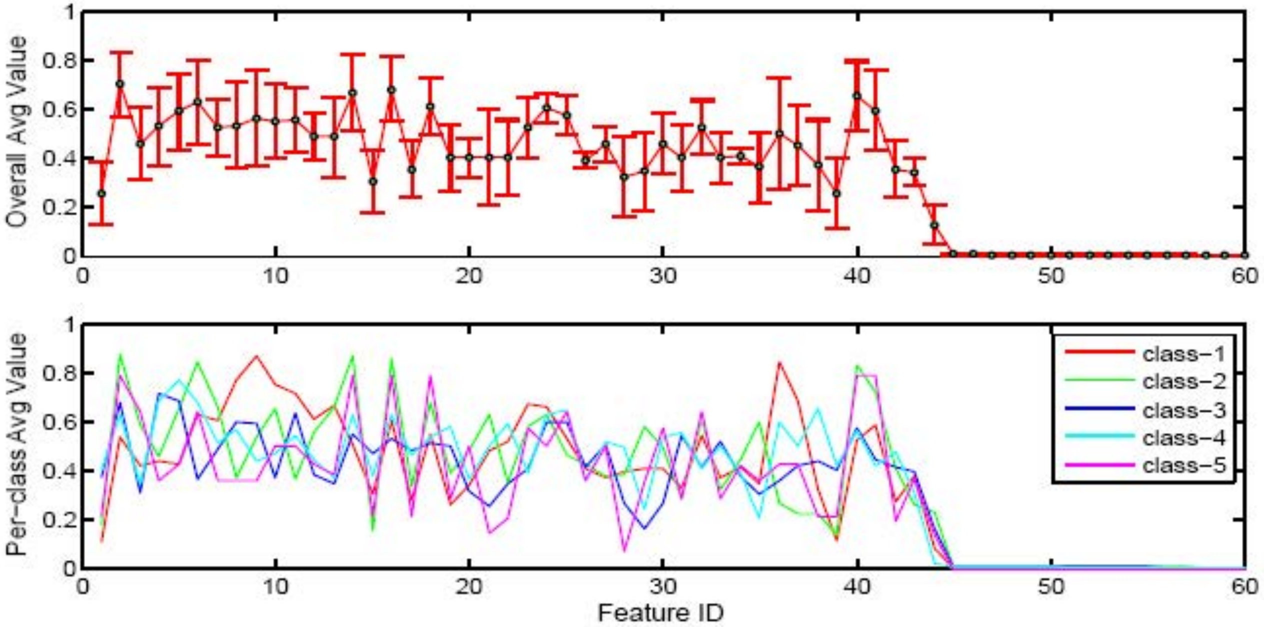
- Posterior constraints with max-margin principle

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

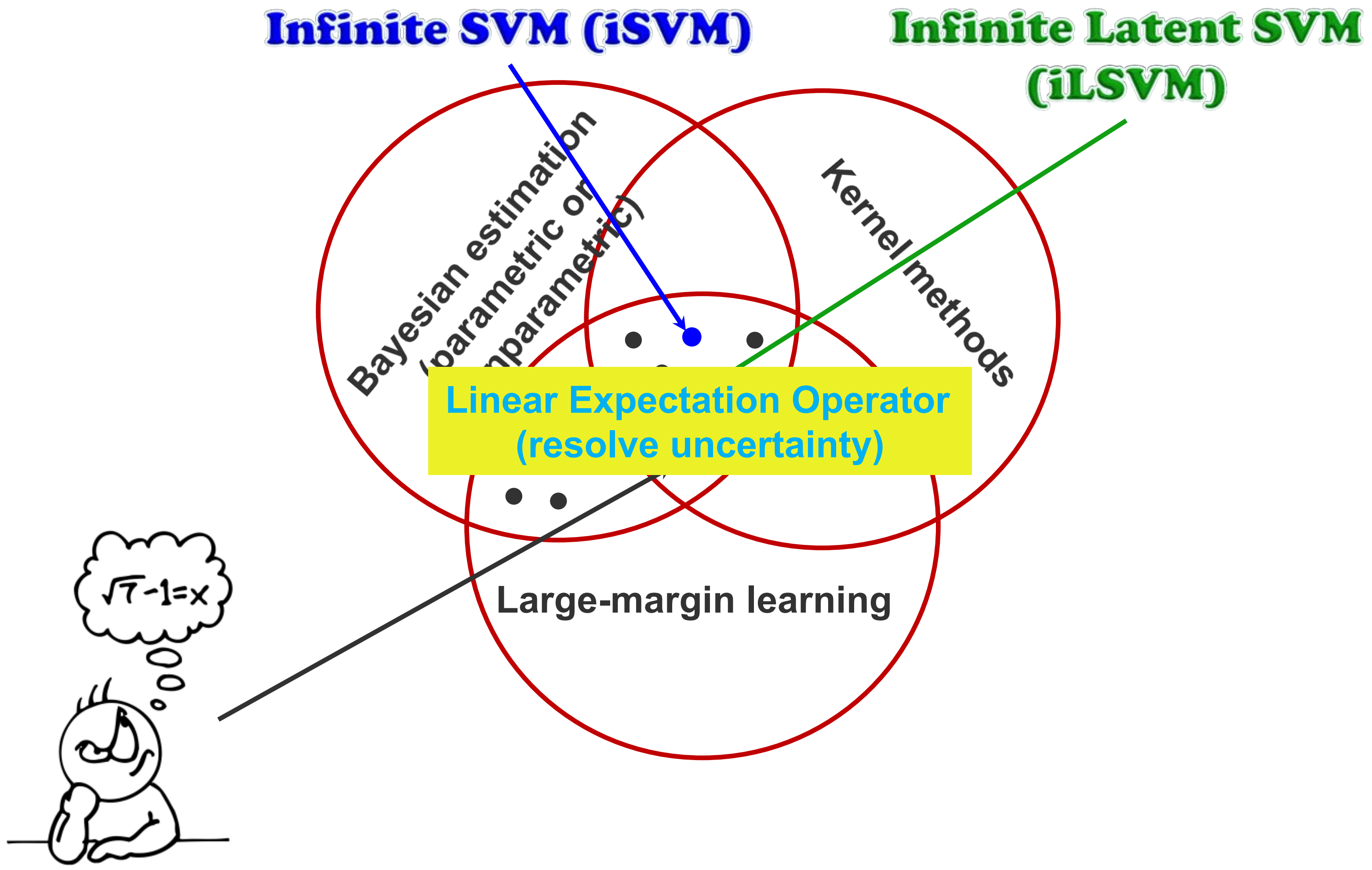
Experimental Results

- 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



Infinite SVM (iSVM)
Infinite Latent SVM (iLSVM)
Bayesian estimation
(parametric or
nparametric)
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.