Bayesian Learning with Rich Side Information

Jun Zhu

dcszj@mail.tsinghua.edu.cn

Department of Computer Science and Technology

Tsinghua University

VALSE 2014, Qingdao

What good for VALSE?

A Probabilistic PC Chair Model

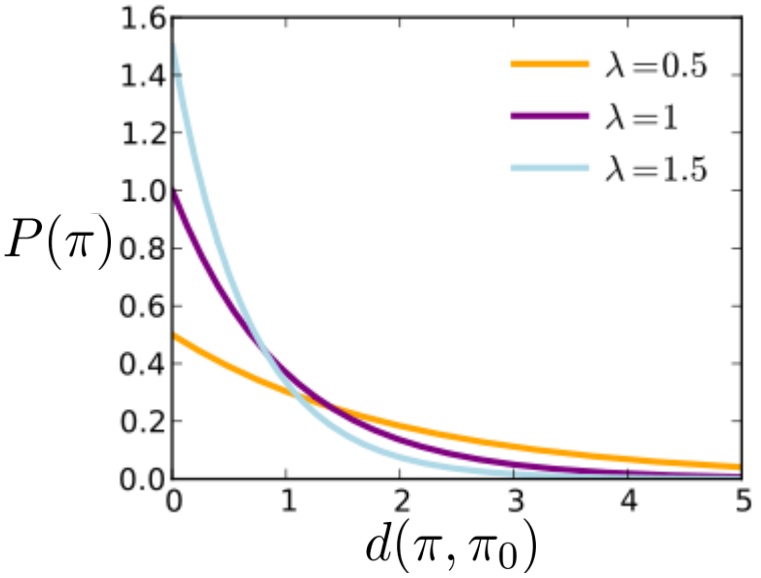
- Arrange a set of invited talks, e.g., A, B, C, D
  - A uniform permutation model

$$P([A,C,B,D]) = P([A,D,C,B]) = \cdots = \frac{1}{4!}$$

A Probabilistic PC Chair Model

- Arrange a set of invited talks
  - With a preferred list
    - PC chairs offer a concentration center $\pi_0 = [C, B, A, D]$
    - A generalized Mallows model is defined

$$P(\pi) \propto \exp\Big(-\lambda d(\pi, \pi_0)\Big)$$



A Probabilistic PC Chair Model

- Arrange a set of invited talks
  - Prior knowledge
    - conjugate prior exists for generalized Mallows models
  - Bayesian updates can be done with Bayes' rule

A Probabilistic PC Chair Model

- Arrange a set of invited talks
  - Side constraints
    - Mike Jordan can only spend 2 days at ICML
    - Eric Horvitz can only spend 1 day at ICML
    - 院士X必须放在第一天
    - Vision 排在 learning前面
    - ……
- How can we consider them?
  - Lets' do optimization?
  - ***How about if Bayesian?***

What's Bayes & Why be Bayesian?

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

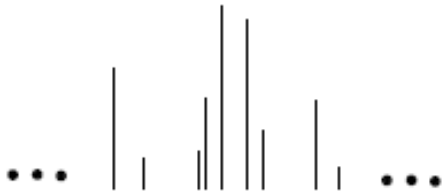
$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}} \propto p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})$$

Examples:

→ see next slide

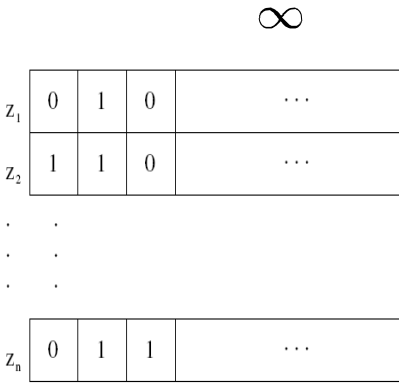
Nonparametric Bayesian Inference

probability measure



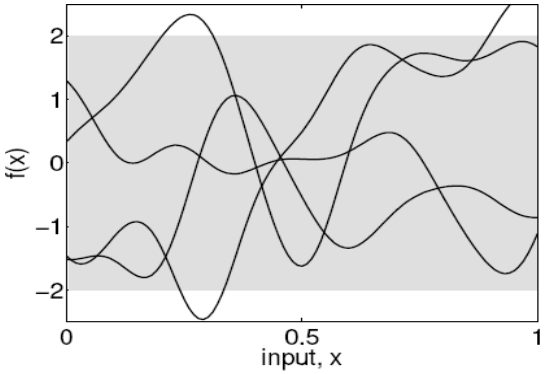
Dirichlet Process Prior [Ferguson, 1973]
+ Multinomial/Gaussian/Softmax likelihood

binary matrix



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

function

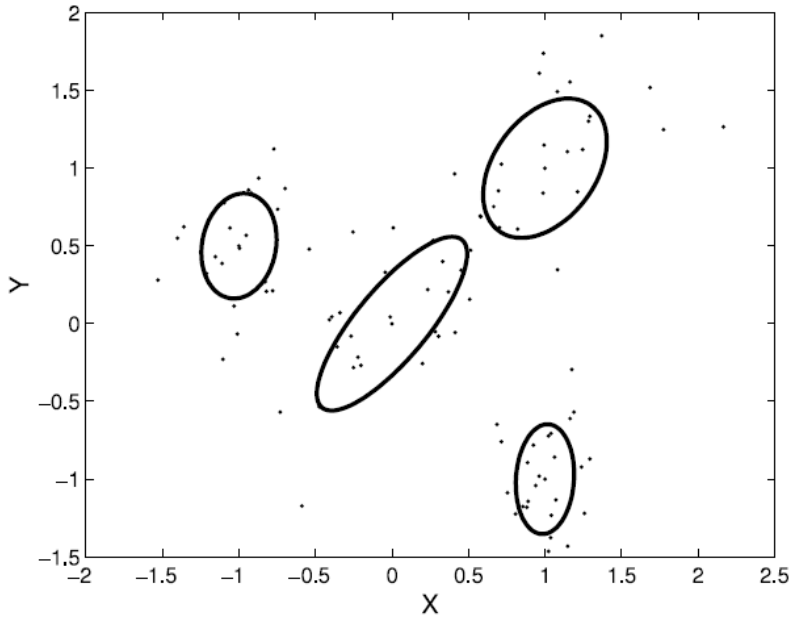


Gaussian Process Prior [Doob, 1944; Rasmussen & Williams, 2006]
+ Gaussian/Sigmoid/Softmax likelihood

Why Bayesian Nonparametrics?

Let the data speak for itself

- 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



Bayesian Inference with Rich Priors

- ***The world is structured and dynamic!***
- Predictor-dependent processes to handle heterogeneous data
  - Dependent Dirichlet Process (MacEachern, 1999)
  - Dependent Indian Buffet Process (Williamson et al., 2010)
  - …
- Correlation structures to relax exchangeability:
  - Processes with hierarchical structures (Teh et al., 2007)
  - Processes with temporal or spatial dependencies (Beal et al., 2002; Blei & Frazier, 2010)
  - Processes with stochastic ordering dependencies (Hoff et al., 2003; Dunson & Peddada, 2007)
  - …

Why be Bayesian?

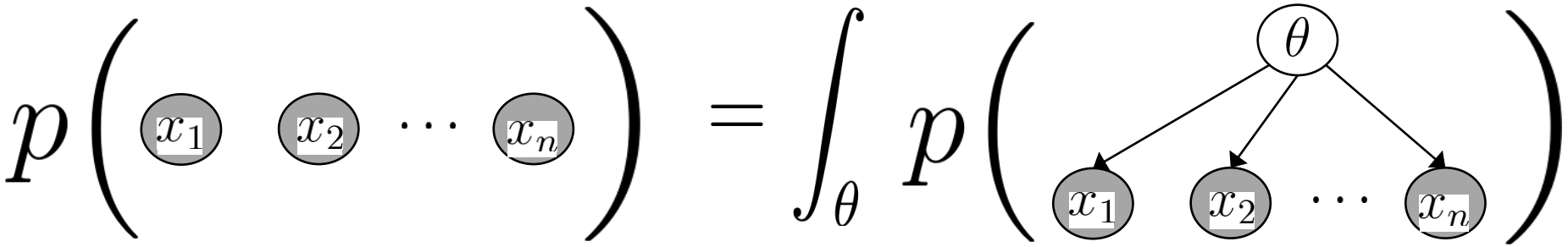
- One of many answers
- Infinite Exchangeability:

$$\forall n,\ \forall \sigma,\ p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$$

- De Finetti's Theorem (1955): if $(x_1, x_2, \ldots)$ are *infinitely exchangeable*, then $\forall n$

$$p(x_1, \ldots, x_n) = \int \Big(\prod_{i=1}^{n} p(x_i|\theta) \Big) dP(\theta)$$

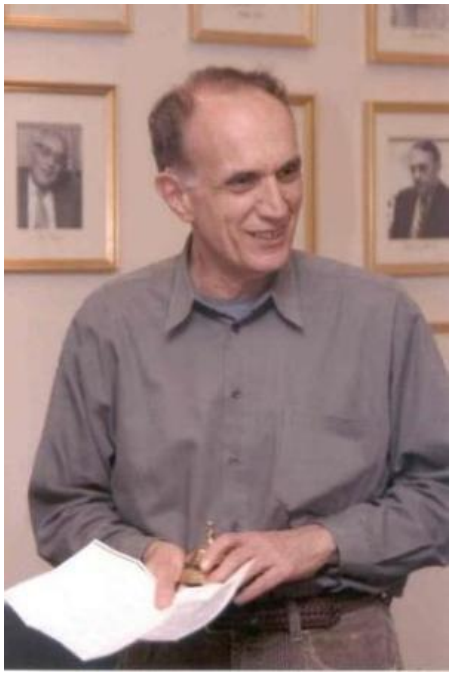
for some random variable θ



Bayes' Theorem in the 21st Century

- 2013 marks the 250th Anniversary of Bayes' theorem

- Bradley Efron, *Science 7 June 2013: Vol. 340 no. 6137 pp. 1177-1178*

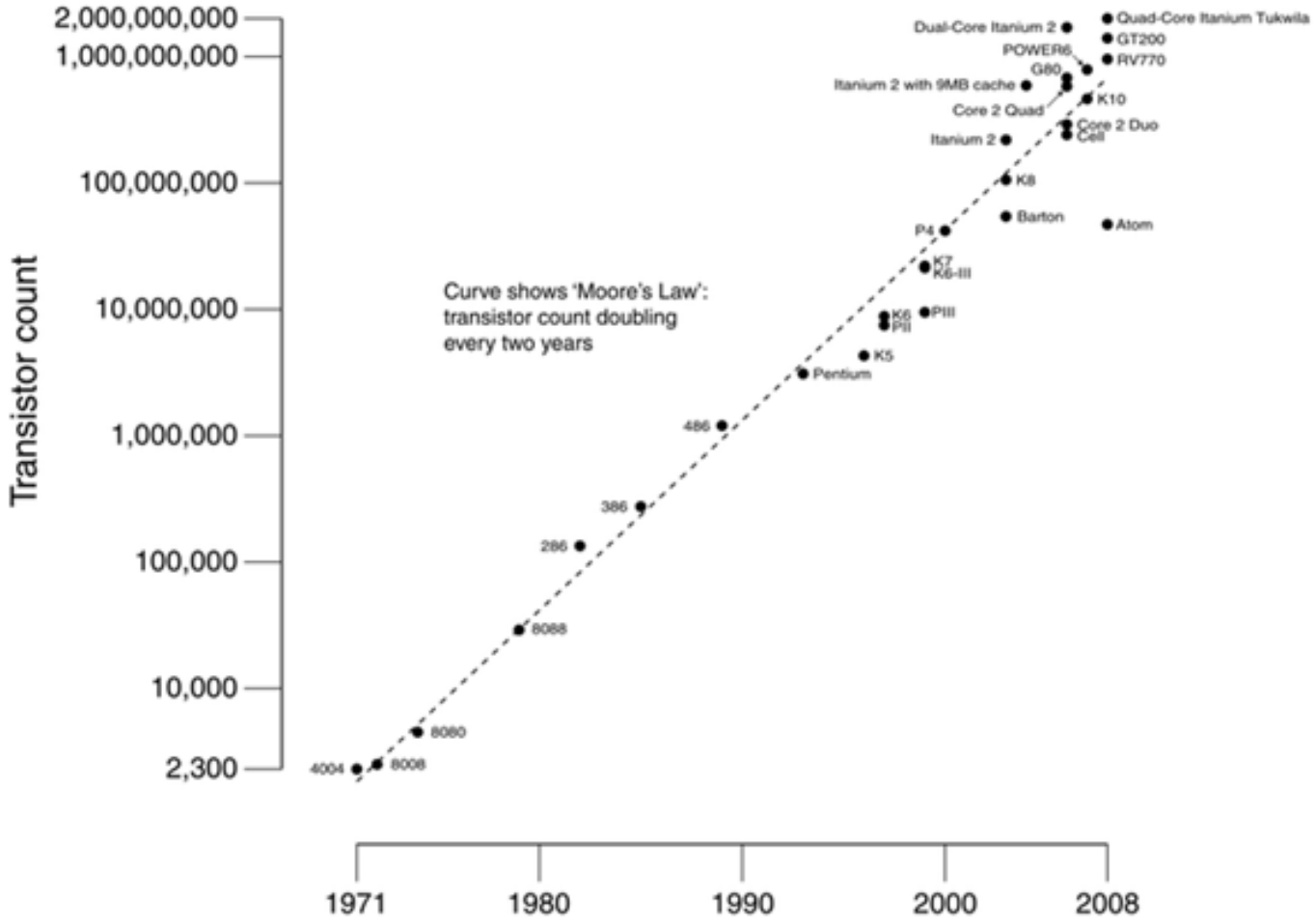


"There are two potent arrows in the statistician's quiver

there is no need to go hunting armed with only one."

Overfitting in Big Data

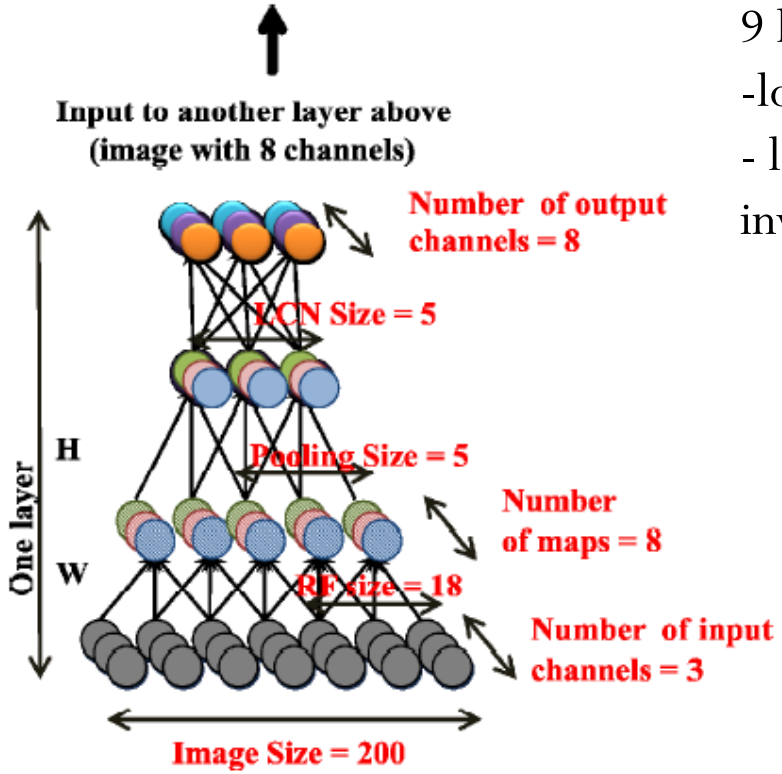
- "with more data overfitting is becoming less of a concern"?



Overfitting in Big Data

"Big Model + Big Data + Big/Super Cluster"

Big Learning



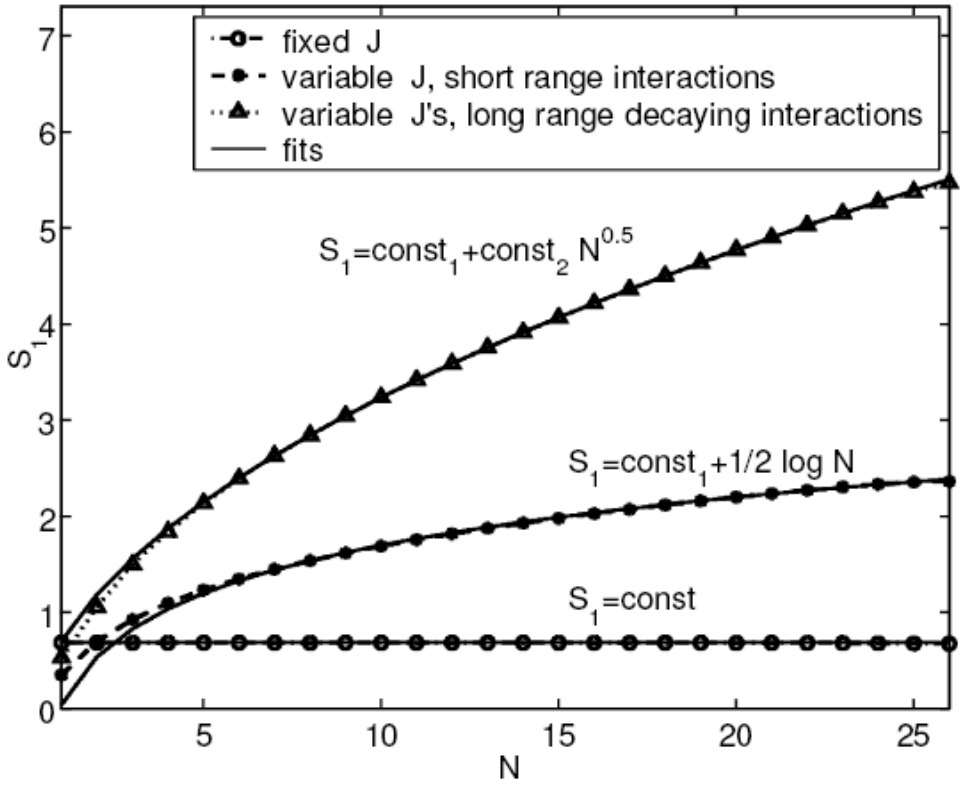
9 layers sparse autoencoder with:
-local receptive fields to scale up;
- local L2 pooling and local contrast normalization for invariant features

- 1B parameters (connections)
- 10M 200x200 images
- train with 1K machines (16K cores) for 3 days

-able to build high-level concepts, e.g., cat faces and human bodies
-15.8% accuracy in recognizing 22K objects (70% relative improvements)

Overfitting in Big Data

- **Predictive information** grows slower than the amount of Shannon entropy (Bialek et al., 2001)



Overfitting in Big Data

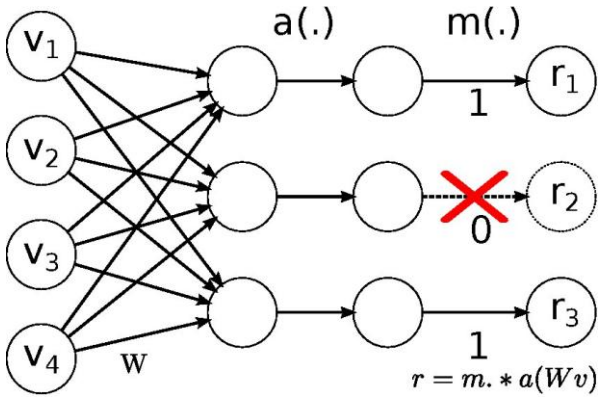
- **Predictive information** grows slower than the amount of Shannon entropy (Bialek et al., 2001)

Model capacity grows faster than the amount of predictive information!

Overfitting in Big Data

- Surprisingly, regularization to prevent overfitting is ***increasingly important***, rather than increasingly irrelevant!

- Increasing research attention, e.g., dropout training (Hinton, 2012)



- More theoretical understanding and extensions
  - MCF (van der Maaten et al., 2013); Logistic-loss (Wager et al., 2013); Dropout SVM (Chen, Zhu et al., 2014)

Therefore …

- Computationally efficient Bayesian models are becoming increasingly relevant in Big data era
  - **Relevant**: high capacity models need a protection
  - **Efficient**: need to deal with large data volumes

Challenges of Bayesian Inference

Building an *Automated* Statistician

- Modeling
  - scientific and engineering data
  - rich side information
- Inference/learning
  - discriminative learning
  - large-scale inference algorithms for Big Data
- Applications
  - social media
  - adaptation

Regularized Bayesian Inference

Regularized Bayesian Inference?

posterior — likelihood model — prior

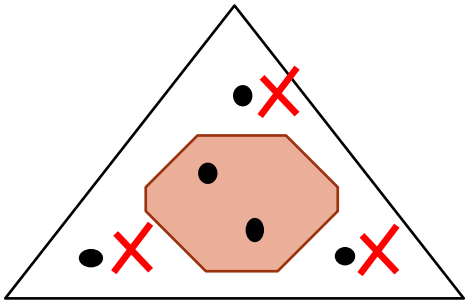
$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}}$$

- How to consider side constraints?

Not obvious!

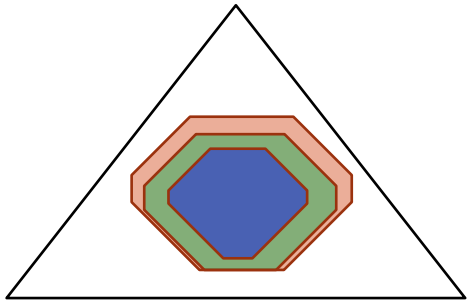
hard constraints

(A single feasible space)



soft constraints

(many feasible subspaces with different complexities/penalties)



Bayesian Inference as an Opt. Problem

Wisdom never forgets that all things have two sides

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

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

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

direct but trivial constraints on posterior distribution

Regularized Bayesian Inference

Constraints can encode rich structures/knowledge

- **Bayesian inference with posterior regularization:**

'unconstrained' equivalence:

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

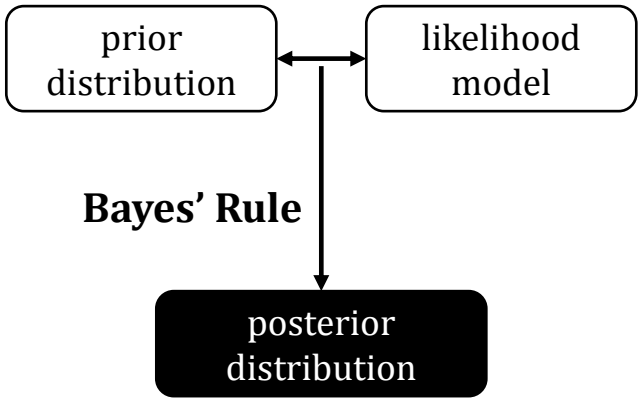
posterior regularization

- Consider both hard and soft constraints
- Convex optimization problem with nice properties
- Can be effectively solved with convex duality theory

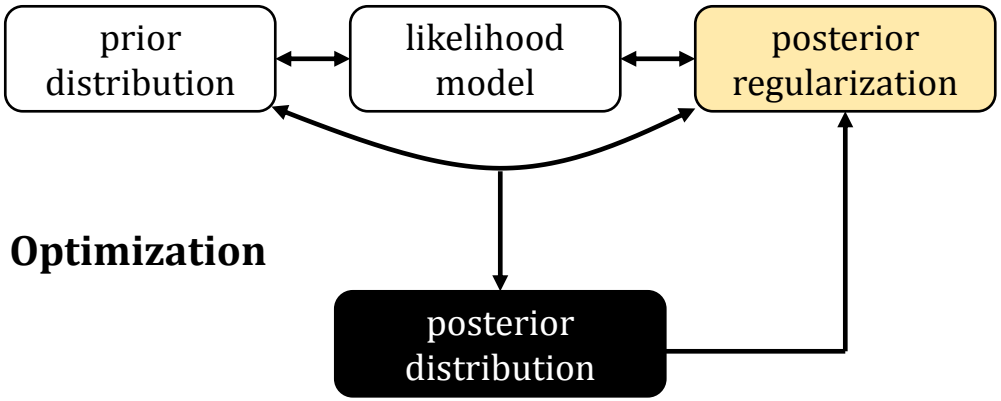
[Zhu, Chen, & Xing, JMLR, in press, 2014]

A High-Level Comparison

Bayes:



RegBayes:



More Properties

- Representation Theorem:
  - the optimum distribution is:

$$\hat{q}_{\hat{\phi}}(\mathbf{M}) = p(\mathbf{M}, \mathcal{D}) \exp(\langle \hat{\phi}, \psi(\mathbf{M}; \mathcal{D}) \rangle - \Lambda_{\hat{\phi}})$$

  - where $\hat{\phi}$ is the solution of the convex dual problem

- Putting constraints on priors is a special case
  - constraints on priors are special cases of posterior regularization

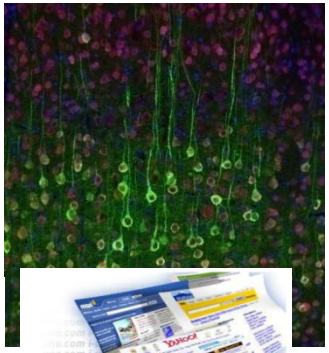
- RegBayes is more flexible than Bayes' rule
  - exist some RegBayes distribution: no implicit prior and likelihood that give back it by Bayes' rule

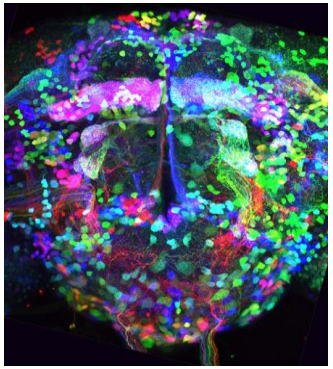
[Zhu, Chen, & Xing, JMLR, in press, 2014]

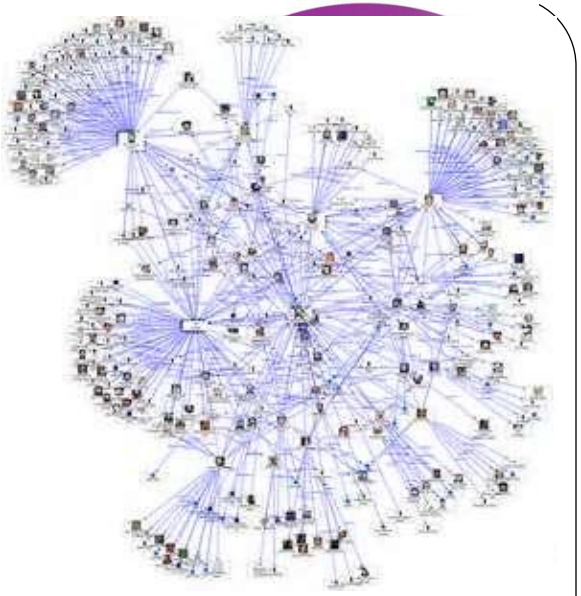
Ways to Derive Posterior Regularization

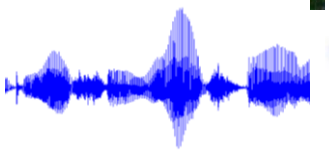
- **From learning objectives**
  - Performance of posterior distribution can be evaluated when applying it to a learning task
  - Learning objective can be formulated as Pos. Reg.
- **From domain knowledge**
  - Elicit expert knowledge
  - E.g., first-order logic rules
- **Others …**
  - E.g., decision making, cognitive constraints, etc.

Adaptive, Discriminative, Scalable Representation Learning





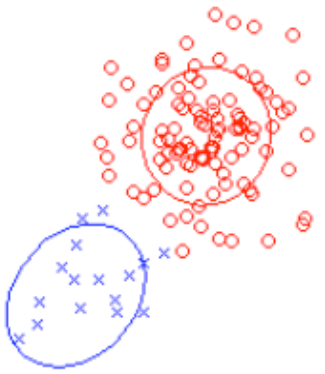


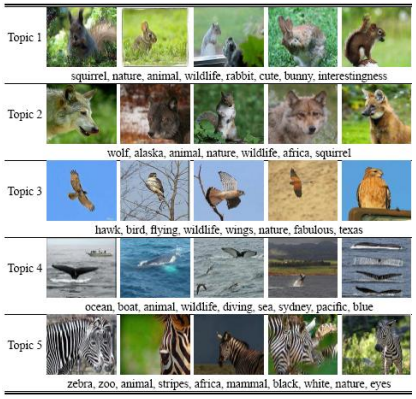


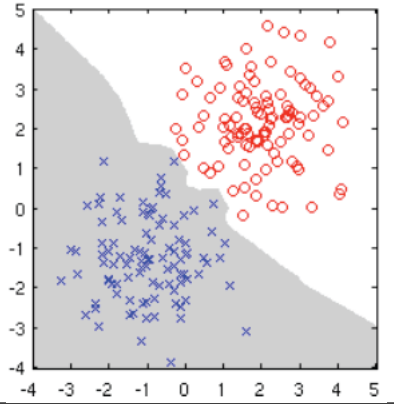
Input Data

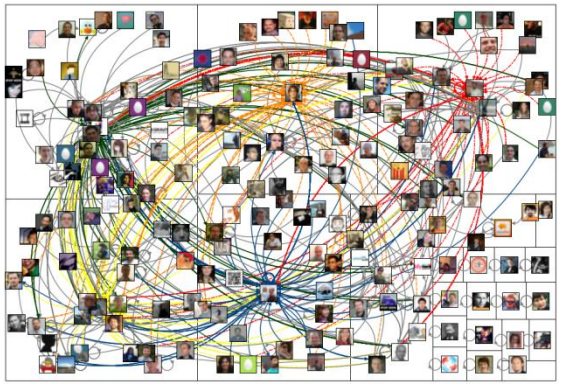
Ideal paradigm that computers help solve big data problems

Inference, Decision, Reasoning







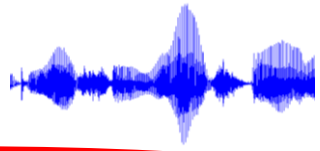


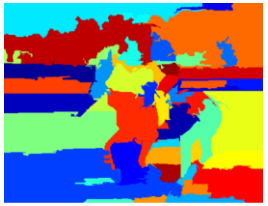
A Conventional Data Analysis Pipeline

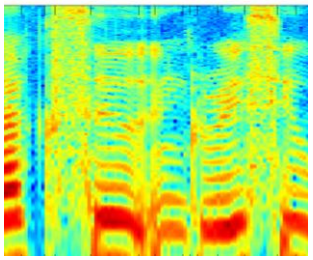
Input Data

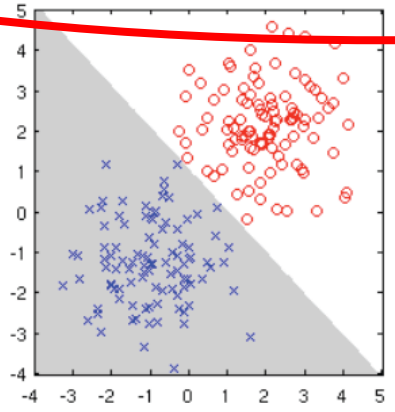
Feature Representation

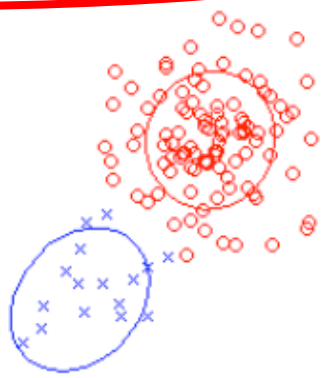
Computer Algorithms

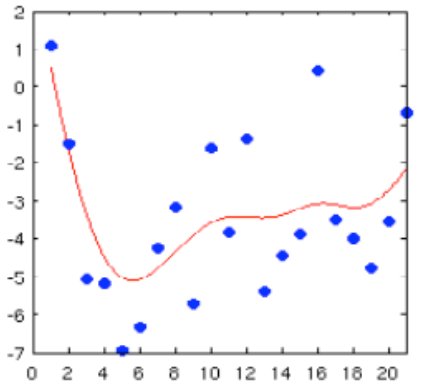




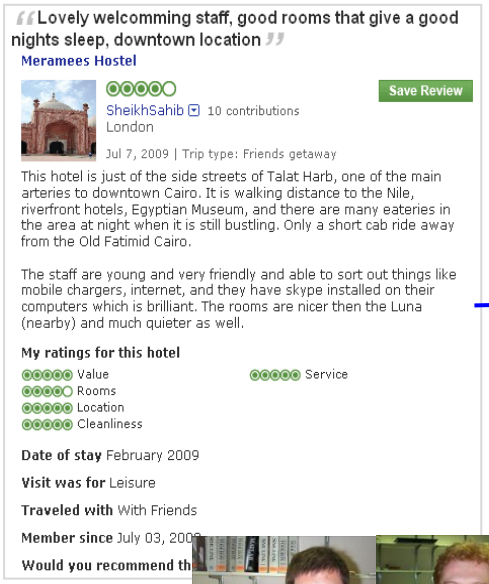








Representation Learning

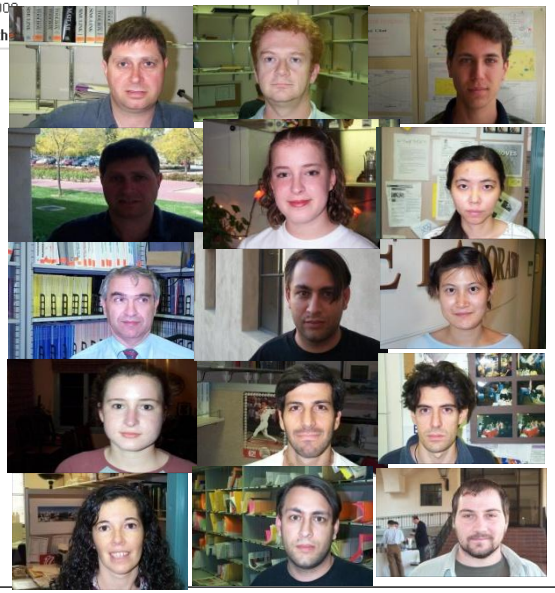


Learning Algorithms

E.g., Topic Models

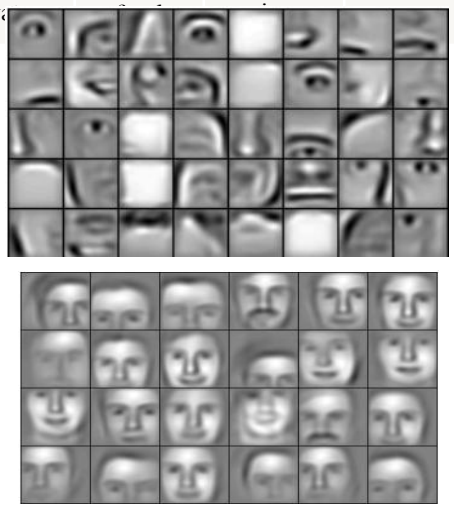
Axis's of a semantic representation space:

T1	T2	T3	T4	T5	T6	T7
told	place	hotel	hotel	beach	beach	great
dirty	hotel	food	area	pool	resort	good
room	room	bar	staff	resort	pool	nice
front	days	day	pool	food	ocean	lovely
asked	time	pool	breakfast	island	island	beautiful
hotel	day	time	day	kids	kids	excellent
bad	night	service	view	trip	good	wonderful
small	people	holiday	location	service	restaurants	comfortable
worst	stay	room	service	day	enjoyed	beach
poor	water	people	walk	staff	loved	friendly
called	rooms	night	time	time	trip	fresh
rude	food	wa				amazing



Learning Algorithms

E.g., Deep Networks



[Figures from (Lee et al., ICML2009)]

Some Key Issues

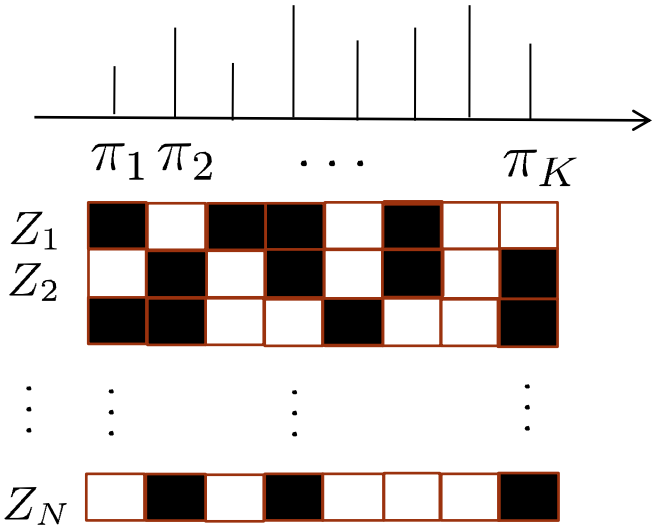
- Discriminative Ability
  - Are the representations good at solving a task, e.g., distinguishing different concepts?
  - Can they generalize well to unseen data?
  - Can the learning process effectively incorporates domain knowledge?
- Model Complexity
  - How many dimensions are sufficient to fit a given data set?
  - Can the models adapt when environments change?
- Sparsity/Interpretability
  - Are the representations compact or easy to interpret?
- Scalability
  - Can the algorithms scale up to Big Data applications?

Bayesian Latent Feature Models (finite)

- A random finite binary latent feature models

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

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



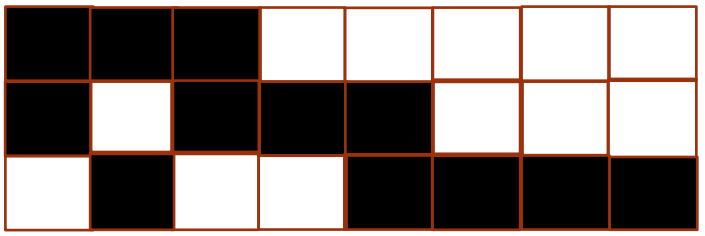
- π_k is the relative probability of each feature being on

- giving the latent structure that's used to generate the data, e.g.,

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

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 $K_i \sim \text{Poisson}(\frac{\alpha}{i})$



cust 1: new dishes 1-3

cust 2: old dishes 1,3; new dishes 4-5

cust 3: old dishes 2,5; new dishes 6-8

$$Z \sim \mathcal{IBP}(\alpha)$$

[Griffiths & Ghahramani, NIPS 2005]

Posterior Constraints – classification

- Suppose latent features $\mathbf{z}$ are given, we define *latent discriminant function*:

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

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

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

- Posterior constraints with max-margin principle

$$\forall n \in \mathcal{I}_{\text{tr}} : y_n f(\mathbf{x}_n; p(\mathbf{Z}, \boldsymbol{\eta})) \geq 1 - \xi_n$$

- Convex U function

$$U(\xi) = C \sum_{n \in \mathcal{I}_{\text{tr}}} \xi_n$$

The RegBayes Problem

$$\min_{q(\mathbf{Z},\mathbf{W},\boldsymbol{\eta})} \quad \mathcal{L}(q(\mathbf{Z},\mathbf{W},\boldsymbol{\eta}) + 2c \cdot \mathcal{R}(q(\mathbf{Z},\mathbf{W},\boldsymbol{\eta}))$$

- where $\mathcal{L}(q) = \mathrm{KL}(q \| \pi(\mathbf{Z},\mathbf{W},\eta)) - \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{Z},\mathbf{W})]$
- the hinge loss (posterior regularization) is

$$\mathcal{R}(q) = \sum_n \max(0, 1 - y_n f(\mathbf{x}_n; q(\mathbf{Z},\boldsymbol{\eta}))$$

Posterior Regularization with a Gibbs Classifier

- Posterior distribution to learn

$$q(\mathbf{Z}, \boldsymbol{\eta})$$

- Gibbs classifier randomly draws a sample to make prediction

$$(\mathbf{Z}, \boldsymbol{\eta}) \sim q(\mathbf{Z}, \boldsymbol{\eta})$$

  - For classification, we measure the loss of classifier $(\mathbf{Z}, \boldsymbol{\eta})$

$$\mathcal{R}(\mathbf{Z}, \boldsymbol{\eta}) = \sum_n \max(0, 1 - y_n f(\mathbf{x}_n; \mathbf{Z}, \boldsymbol{\eta})$$

  - It minimizes the expected loss

$$\mathcal{R}'(q) = \mathbb{E}_q \left[\sum_n \max(0, 1 - y_n f(\mathbf{x}_n; \mathbf{Z}, \boldsymbol{\eta}) \right]$$

Comparison

- Expected hinge-loss is an upper bound

$$\mathcal{R}'(q) \geq \mathcal{R}(q)$$

- For averaging classifier, the RegBayes problem is suitable for variational inference with truncation (Zhu et al., JMLR 2014)

- For Gibbs classifier, the RegBayes problem is suitable for MCMC without truncation

More Details on MCMC

- RegBayes problem

$$\min_{q(\mathbf{Z},\mathbf{W},\boldsymbol{\eta})} \quad \mathcal{L}(q(\mathbf{Z},\mathbf{W},\boldsymbol{\eta}) + 2c \cdot \mathcal{R}'(q(\mathbf{Z},\mathbf{W},\boldsymbol{\eta}))$$

- The solution is

$$q(\mathbf{Z},\mathbf{W},\boldsymbol{\eta}) = \frac{\pi(\mathbf{Z},\mathbf{W},\boldsymbol{\eta})p(\mathbf{X}|\mathbf{Z},\mathbf{W})\phi(\mathbf{y}|\mathbf{Z},\boldsymbol{\eta})}{\psi(\mathbf{X},\mathbf{y})}$$

  - where

$$\phi(\mathbf{y}|\mathbf{Z},\boldsymbol{\eta}) = \prod_n \phi(y_n|\mathbf{Z},\boldsymbol{\eta}) = \prod_n \exp\{-2c\max(0, 1 - y_n f(\mathbf{x}_n;\mathbf{Z},\boldsymbol{\eta})\}$$

More Details on MCMC

- Scale mixture representation: $\zeta_n = 1 - y_n f(\mathbf{x}_n; \mathbf{Z}, \boldsymbol{\eta})$

$$\phi(y_n|\mathbf{Z}, \eta) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_n}} \exp\left(-\frac{(\lambda_n + c\zeta_n)^2}{2\lambda_n}\right) d\lambda_n$$

  - follows (Polson & Scott, 2011)

- Data augmentation representation

$$q(\mathbf{Z}, \mathbf{W}, \eta) = \int q(\mathbf{Z}, \mathbf{W}, \eta, \lambda) d\lambda$$

$$\text{where } q(\mathbf{Z}, \mathbf{W}, \eta, \lambda) = \frac{\pi(\mathbf{Z}, \mathbf{W}, \eta) p(\mathbf{X}|\mathbf{Z}, \mathbf{W}) \phi(\mathbf{y}, \lambda|\mathbf{Z}, \eta)}{\psi(\mathbf{X}, \mathbf{y})}$$

$$\phi(\mathbf{y}, \lambda|\mathbf{Z}, \eta) = \prod_n \frac{1}{\sqrt{2\pi\lambda_n}} \exp\left(-\frac{(\lambda_n + c\zeta_n)^2}{2\lambda_n}\right)$$

More Details on MCMC

- Data augmented posterior

$$q(\mathbf{Z}, \mathbf{W}, \eta, \lambda) = \frac{\pi(\mathbf{Z}, \mathbf{W}, \eta) p(\mathbf{X}|\mathbf{Z}, \mathbf{W}) \phi(\mathbf{y}, \lambda|\mathbf{Z}, \eta)}{\psi(\mathbf{X}, \mathbf{y})}$$

- A Gibbs sampler is as follows
  - Sample $\boldsymbol{\eta} \sim q(\boldsymbol{\eta}|\mathbf{Z}, \mathbf{W}, \lambda) \propto \pi(\boldsymbol{\eta}) \phi(\mathbf{y}, \lambda|\mathbf{Z}, \boldsymbol{\eta})$
    - a Gaussian distribution if the prior is Gaussian
  - Sample $\lambda \sim q(\lambda|\mathbf{Z}, \mathbf{W}, \eta) \propto \phi(\mathbf{y}, \lambda|\mathbf{Z}, \eta)$
    - a generalized inverse Gaussian distribution
  - Sample $(\mathbf{Z}, \mathbf{W}) \sim q(\mathbf{Z}, \mathbf{W}|\eta, \lambda) \propto \pi(\mathbf{Z}, \mathbf{W}) p(\mathbf{X}|\mathbf{Z}, \mathbf{W}) \phi(\mathbf{y}, \lambda|\mathbf{Z}, \eta)$
    - Similar as the Gaussian infinite latent feature model (Griffiths & Ghahramani, 2005)

PAC-Bayes Theory

- Theorem (Germain et al., 2009):
  - for any distribution D ; for any set $\mathcal{H}$ of classifiers, for any prior P , for any convex function

$$\phi : \ [0,1] \times [0,1] \to \mathbb{R}$$

  - for any posterior Q , for any $\delta \in (0,1]$, the following inequality holds with a high probability ($\geq 1-\delta$)

$$\phi\left(R_S(G_Q), R(G_Q)\right) \leq \frac{1}{N}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{C(N)}{\delta}\right)\right]$$

  - where $C(N) = \mathbb{E}_{S\sim D^N}\mathbb{E}_{h\sim P}\left[e^{N\phi(R_S(h),R(h))}\right]$

RegBayes Classifiers

- PAC-Bayes theory

$$\phi\left(R_S(G_Q), R(G_Q)\right) \leq \frac{1}{N}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{C(N)}{\delta}\right)\right]$$

- RegBayes inference

$$\begin{aligned} \min_{q(\mathcal{H})} \quad & \mathrm{KL}(q(\mathcal{H})\|p(\mathcal{H}|\mathbf{x})) + \Omega(q(\mathcal{H})) \\ \text{s.t.}: \quad & q(\mathcal{H}) \in \mathcal{P}_{\text{prob}}, \end{aligned}$$

- Observations:
  - when the posterior regularization equals to (or upper bounds) the empirical risk

$$\Omega(q(\mathcal{H})) \geq R_S(G_q)$$

  - the RegBayes classifiers tend to have PAC-Bayes guarantees.

Extensions to Multi-task Learning

Multi-task Learning (MTL)

- **[Wikipedia]** MTL is an approach to machine learning that learns a problem together with other related problems, using a ***shared representation***

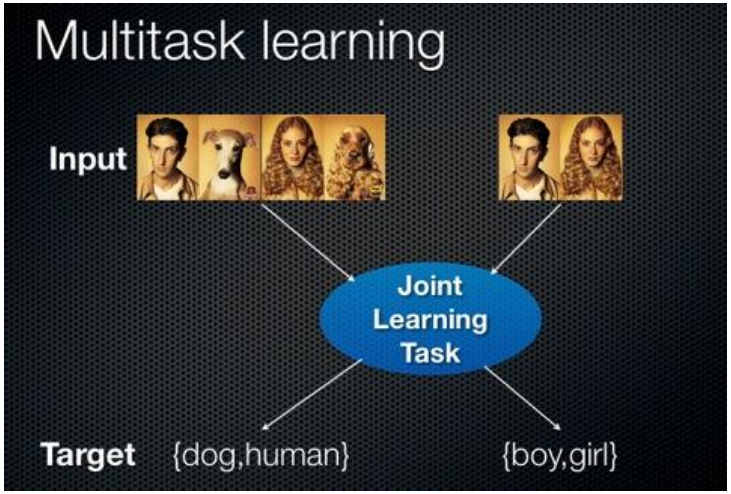


Figure from Wikipedia
Author: Kilian Weinberger

- The goal of MTL is to improve the performance of learning algorithms by learning classifiers for multiple tasks jointly

- It works particularly well if these tasks have some commonality and are generally slightly under sampled

Multi-task Representation Learning

- Assumption:
  - common underlying representation across tasks
- Representative works:
  - ASO (alternating structure optimization): learn a small set of shared features across tasks [Ando & Zhang, 2005]
  - Convex feature learning via sparse norms [Argyriou et al., 2006]

Basic Learning Paradigm

- Tasks: $m = 1, \cdots, M$
- N examples per task

$$(\mathbf{x}_{m1}, y_{m1}), \cdots, (\mathbf{x}_{mN}, y_{mN}) \in \mathbb{R}^D \times \mathbb{R}$$

- Estimate

$$f_m : \mathbb{R}^D \to \mathbb{R}, \ \forall m = 1, \cdots, M$$

- Consider features

$$h_1(\mathbf{x}), \ \cdots, h_K(\mathbf{x})$$

- Predict using functions

$$f_m(\mathbf{x}) = \sum_{k=1}^{K} \eta_{mk} h_k(\mathbf{x})$$

Learning a Projection Matrix

- Tasks: $m = 1, \cdots, M$
- N examples per task

$$(\mathbf{x}_{m1}, y_{m1}), \cdots, (\mathbf{x}_{mN}, y_{mN}) \in \mathbb{R}^D \times \mathbb{R}$$

- Estimate

$$f_m : \mathbb{R}^D \to \mathbb{R}, \ \forall m = 1, \cdots, M$$

- Consider features

$$h_k(\mathbf{x}) = \mathbf{z}_k^\top \mathbf{x}, \ k = 1, \cdots, \infty$$

- Predict using functions ($\mathbf{Z}$ is a $D \times \infty$ projection matrix)

$$f_m(\mathbf{x}; \mathbf{Z}, \boldsymbol{\eta}) = \sum_{k=1}^{\infty} \eta_{mk} (\mathbf{z}_k^\top \mathbf{x}) = \boldsymbol{\eta}_m^\top (\mathbf{Z}^\top \mathbf{x})$$

Max-margin Posterior Regularizations

- Similar as in infinite latent SVMs
  - Averaging classifier

$$y_{mn}\mathbb{E}_q[f_m(\mathbf{x}_{mn};\mathbf{Z},\boldsymbol{\eta})] \geq 1-\xi_{mn}$$

    - The hinge loss

$$\mathcal{R} = \sum_{m,n\in\mathcal{I}_{\text{tr}}^m} \max\left(0, 1-y_{mn}\mathbb{E}_q[f_m(\mathbf{x}_{mn};\mathbf{Z}\boldsymbol{\eta})]\right)$$

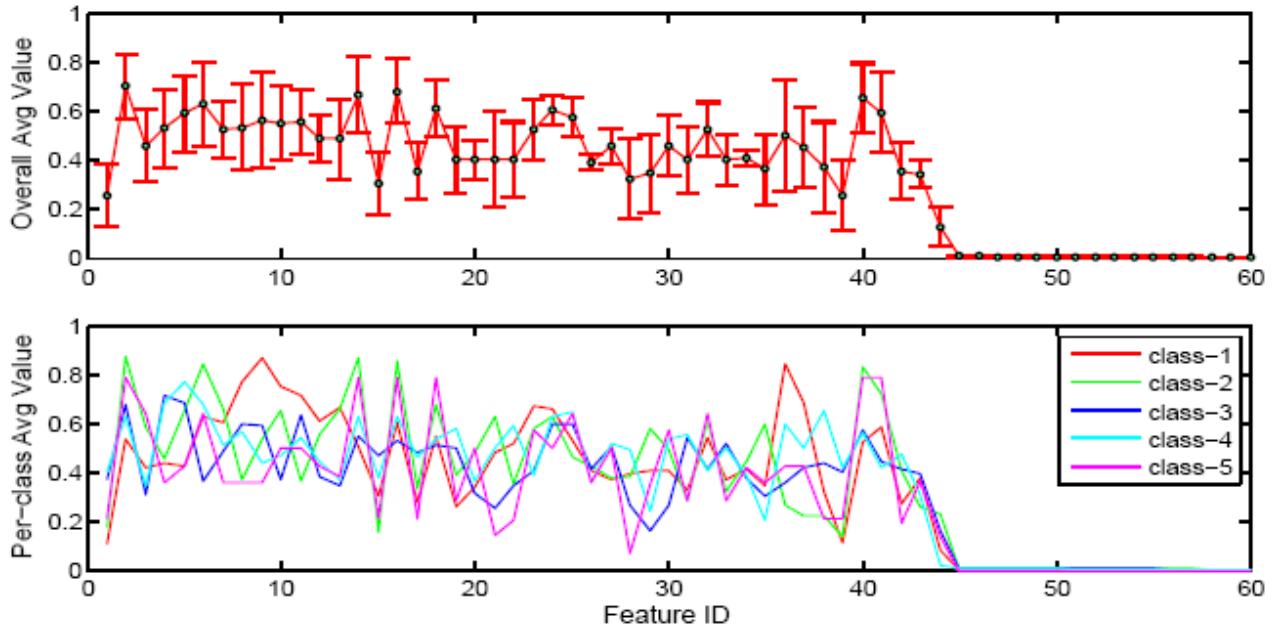
  - Gibbs classifier

$$\mathcal{R}' = \mathbb{E}_q\left[\sum_{m,n\in\mathcal{I}_{\text{tr}}^m} \max\left(0, 1-y_{mn}f_m(\mathbf{x}_{mn};\mathbf{Z}\boldsymbol{\eta})\right)\right]$$

Experimental Results

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

Model	TRECVID2003		Flickr	
	Accuracy	F1 score	Accuracy	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



Experimental Results

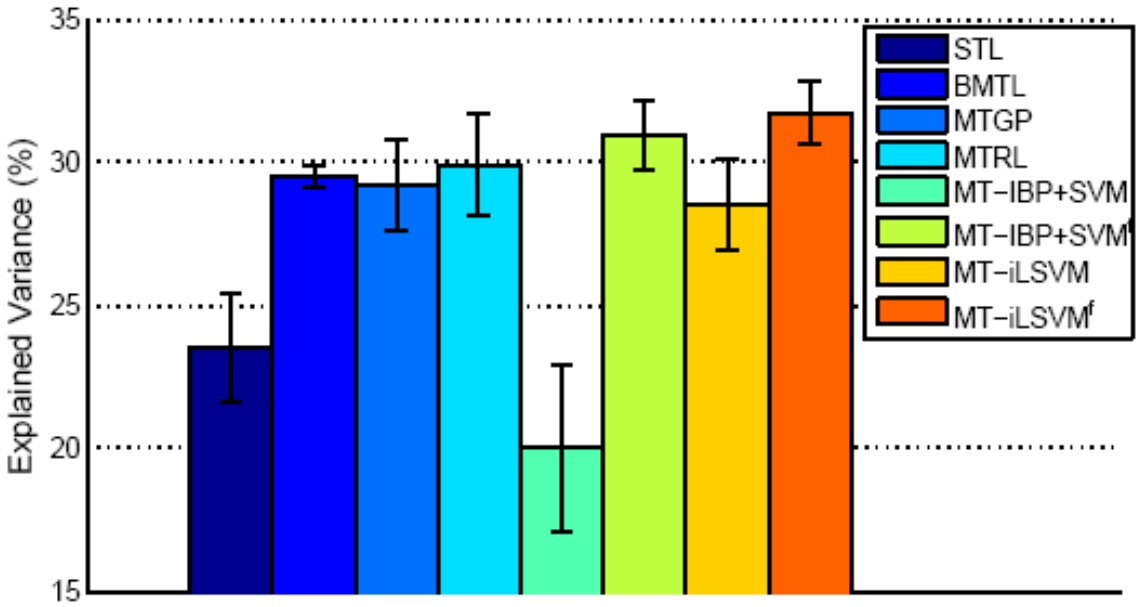
- Multi-label Classification (multiple binary classification)
  - Accuracy and F1 scores (Micro & Macro) on Yeast and Scene datasets

Model	Acc	F1-Macro	F1-Micro
YaXue [Xue et al., 2007]	0.5106	0.3897	0.4022
Piyushrai [Piyushrai et al., 2010]	0.5424	0.3946	0.4112
MT-iLSVM	0.5792 ± 0.003	0.4258 ± 0.005	0.4742 ± 0.008
Gibbs MT-iLSVM	0.5851 ± 0.005	0.4294 ± 0.005	0.4763 ± 0.006

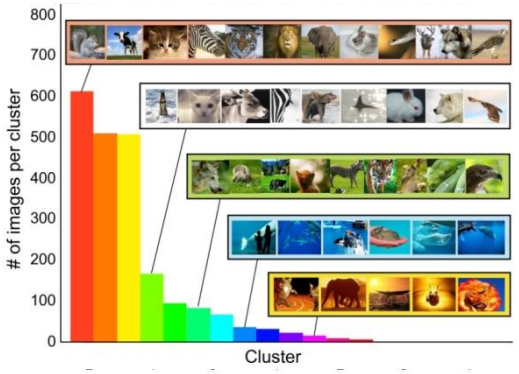
Model	Acc	F1-Macro	F1-Micro
YaXue [Xue et al., 2007]	0.7765	0.2669	0.2816
Piyushrai [Piyushrai et al., 2010]	0.7911	0.3214	0.3226
MT-iLSVM	0.8752 ± 0.004	0.5834 ± 0.026	0.6148 ± 0.020
Gibbs MT-iLSVM	0.8855 ± 0.004	0.6494 ± 0.011	0.6458 ± 0.011

Experimental Results

- Multi-task Regression
  - School dataset (139 regression tasks) – a standard dataset for evaluating multi-task learning
  - Percentage of *explained variance* (higher, better)



RegBayes with Max-margin Posterior Regularization

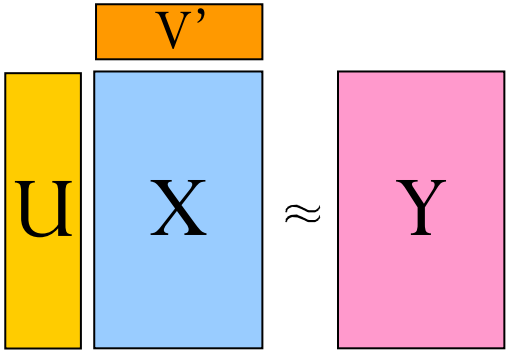


Infinite SVMs

(Zhu et al., ICML'11)

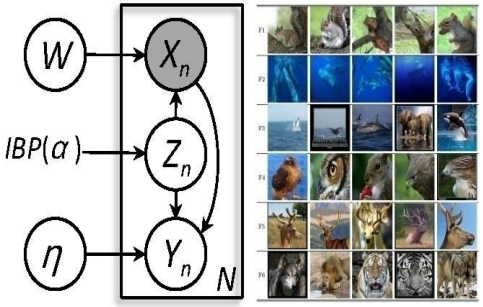
Nonparametric Max-margin Relational Models for Social Link Prediction

(Zhu, ICML'12)



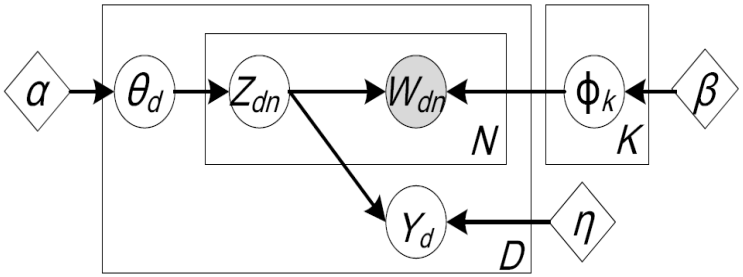
Nonparametric Max-margin Matrix Factorization

(Xu, Zhu, & Zhang, NIPS'12, ICML'13)



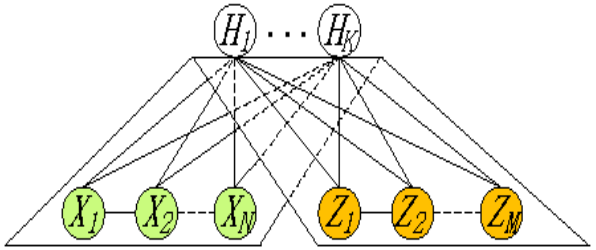
Infinite Latent SVMs

(Zhu, et al., JMLR'14)



Max-margin Topics and Fast Inference

(Zhu, et al., JMLR'12; Zhu et al., JMLR'14)



Multimodal Representation Learning

(Chen, Zhu, et al, PAMI'12)

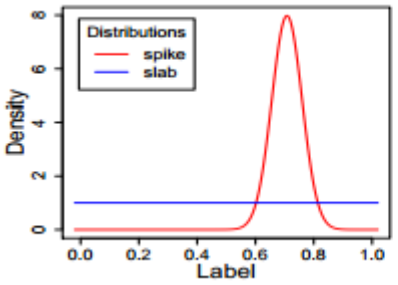
* Works from other groups are not included.

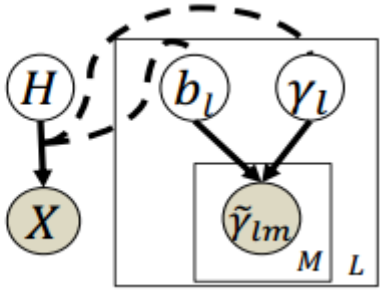
Robust RegBayes with FOL Knowledge

- Goal of "Statistics + Knowledge"

- Incorporate domain knowledge into Bayes models
  - elicit informative priors
  - RegBayes: posterior regularization via FOL rules

- Resolve the uncertainty of domain knowledge in FOL forms
  - A selective spike-and-slab prior





[Mei, Zhu & Zhu, ICML 2014]

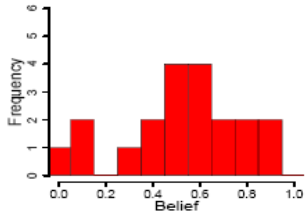
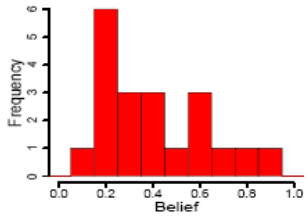
Some Empirical Results

	Test Set Perplexity		
	LDA	Fold·all	LogicLDA
COMP	1531 ± 12	1537 ± 11	$\mathbf{1463} \pm 5$
CON	$\mathbf{1206} \pm 6$	1535 ± 10	$\mathbf{1216} \pm 11$
POL	3218 ± 13	3220 ± 13	$\mathbf{3176} \pm 12$
HDG	940 ± 6	973 ± 7	$\mathbf{885} \pm 2$

Proportion of Satisfied Logic Rules		
LDA	Fold·all	LogicLDA
0.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.01
0.07 ± 0.04	0.67 ± 0.03	0.70 ± 0.00
1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
0.60 ± 0.01	0.95 ± 0.00	0.96 ± 0.01

- LDA: standard unsupervised topic model (Blei et al., 2003)
- Fold.all: hand-tuning rule weights by experts (Andrzejewski et al., 2011)

Robustness to Unreliable Rules

Histogram	Test Set Perplexity		Satisfaction Proportion	
	LogicLDA	RLogicLDA	LogicLDA	RLogicLDA
	1467 ± 6	$\mathbf{1446} \pm 6$	0.39 ± 0.16	0.07 ± 0.06
	$\mathbf{1228} \pm 9$	$\mathbf{1228} \pm 16$	0.49 ± 0.03	0.08 ± 0.03

- More results on interpretability of latent topics and competitive prediction performance (Mei et al., 2014)

Scalable Algorithms

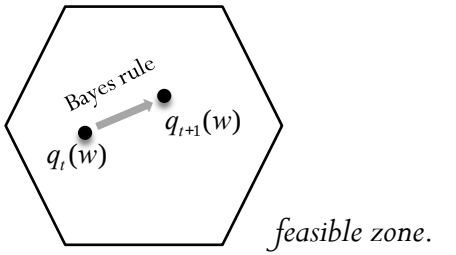
Online Learning Algorithms

- Works on a single machine; Explore **data redundancy**
- Scale up to infinite size data sets, especially for streaming data

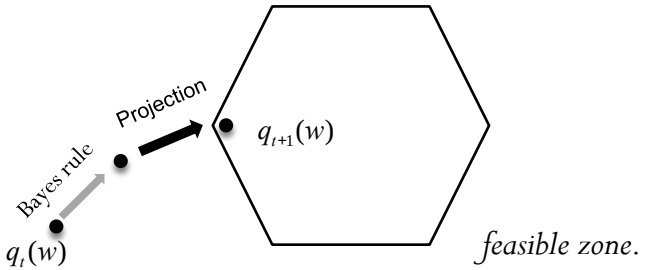
Online Bayesian Passive-Aggressive Learning

$$\min_{q \in \mathcal{F}_t} \mathrm{KL}[q(w) \| q_t(w)] - \mathbb{E}_q[\log p(x_t \mid w)]$$

$$\mathbf{s.t.}: \ell_\epsilon(q(w); x_t, y_t) = 0.$$



(a) Passive update



(b) Aggressive update

Performance is guaranteed with provable regret bounds.

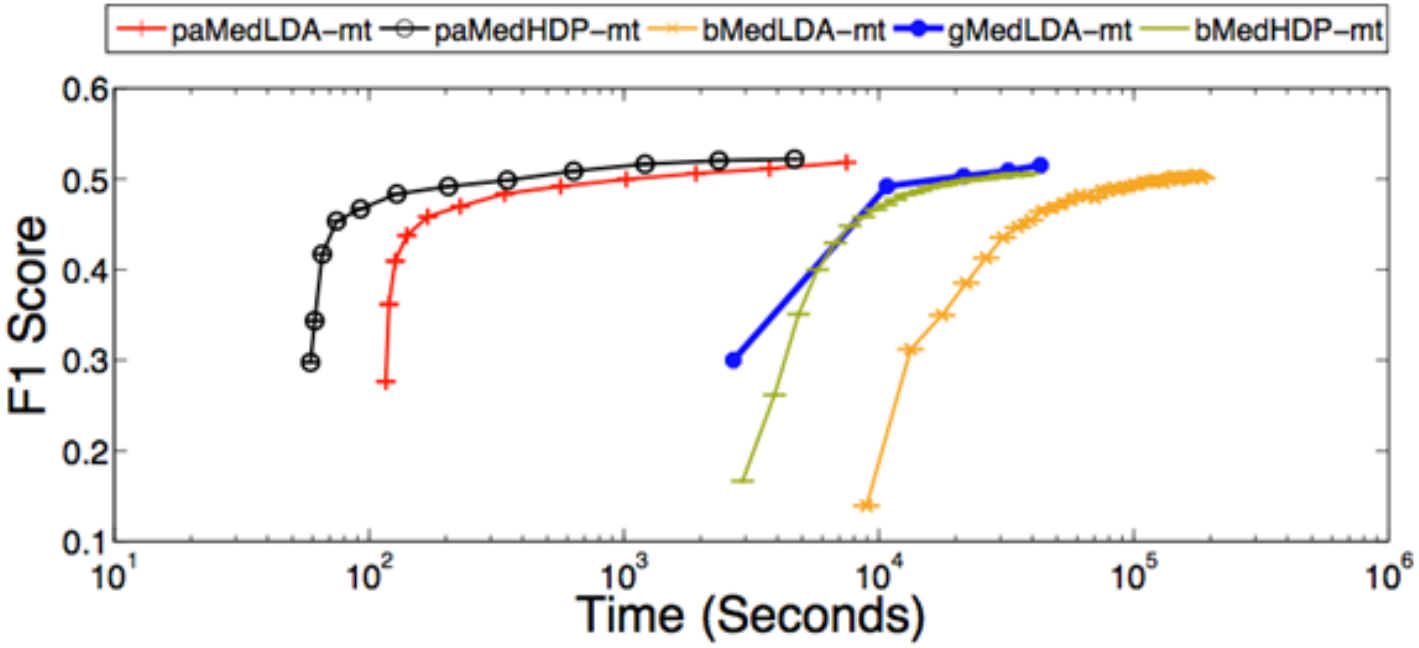
[Shi & Zhu, ICML2014]

Online Learning Algorithms

- Works on a single machine
- Scale up to infinite size data sets, especially for streaming data

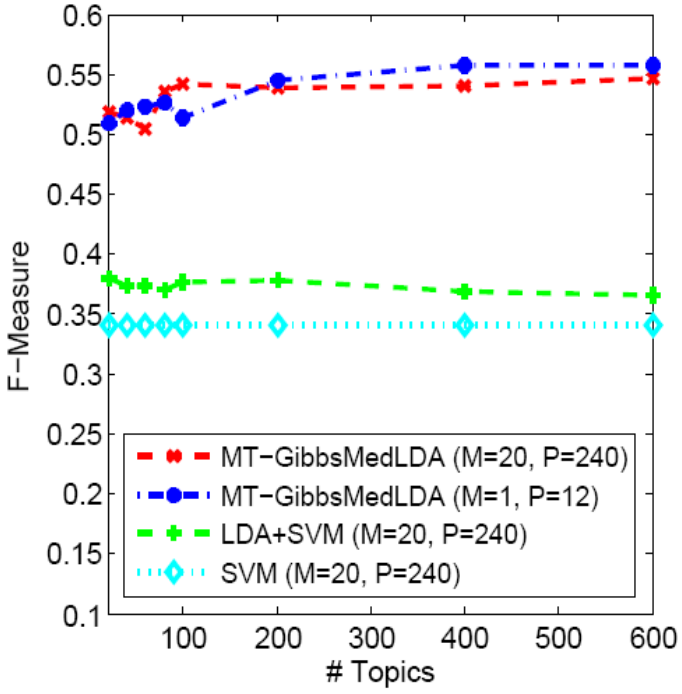
Online Bayesian Passive-Aggressive Learning

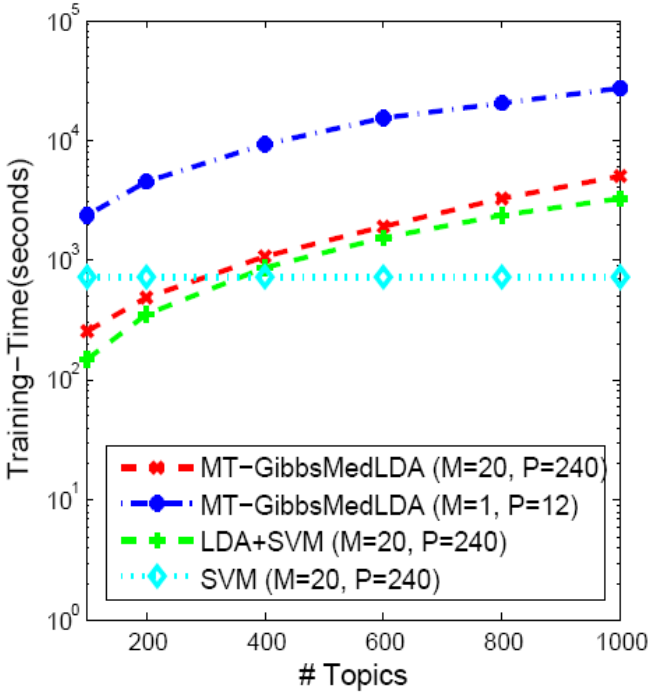
- 1.1M Wiki pages (standard desktop)



Distributed Inference Algorithms

- Leverage big clusters
- Allow learning big models that can't fit on a single machine





- 20 machines;
- 240 CPU cores

- 1.1M multi-labeled Wiki pages
- 20 categories (scale to hundreds/thousands of categories)

[Zhu, Zheng, Zhou, & Zhang, KDD2013]

Summary

- RegBayes:
  - bridge Bayesian methods, learning and optimization
  - offer an extra freedom to incorporate rich side information

- Challenges of Bayesian methods in Big Data
  - effective regularization to avoid overfitting
  - scalable inference algorithms (variational & Monte Carlo)

More on ICML 2014

- Tutorials, Research papers, and Workshops
- 21 June to 26 June, BICC, Beijing

Welcome to join us!

ICML | Beijing

International conference on machine learning, 2014

21-26 JUNE 2014 BEIJING

Acknowledgements

- Collaborators:
  - Prof. Bo Zhang (Tsinghua)、Prof. Eric P. Xing (CMU)、Prof. Li Fei-Fei (Stanford), Xiaojin Zhu (Wisconsin)
  - Amr Ahmed (CMU), Ning Chen (Tsinghua), Ni Lao (CMU), Seunghak Lee (CMU), Li-jia Li (Stanford), Xiaojiang Liu (USTC), Xiaolin Shi (Stanford), Hao Su (Stanford), Yuandong Tian (CMU).
- Students at Tsinghua:
  - Aonan Zhang Minjie Xu Hugh Perkins Bei Chen, Kuan Liu, Shike Mei, Xun Zheng, Jianfei Chen, Wenbo Hu, Yining Wang, Zi Wang, Tianlin Shi, Jingwei Zhuo, Chang Liu, Chao Du, Fei Xia, etc.
- Funding:

Microsoft Research
微软亚洲研究院

Thanks!

Some code available at:

http://www.ml-thu.net/~jun