Laboratory for Statistical Artificial Intelligence & INtegrative Genomics

Recent Advances in Learning SPARSE Structured I/O Models: models, algorithms, and applications

Eric Xing

epxing@cs.cmu.edu

Machine Learning Dept./Language Technology Inst./Computer Science Dept.

Carnegie Mellon University

Structured Prediction Problem

- Unstructured prediction

$$\mathbf{x} = (\, x_{11} \;\; x_{12} \;\; \dots \,) \qquad \mathbf{y} = y_1$$

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \cdots \\ x_{21} & x_{22} & \cdots \\ \vdots & \vdots & \dots \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix}$$

- Structured prediction
  - Part of speech tagging

$\mathbf{x} =$ "Do you want sugar in it?" $\Rightarrow$ $\mathbf{y} =$ <verb pron verb noun prep pron>

  - Image segmentation

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \cdots \\ x_{21} & x_{22} & \cdots \\ \vdots & \vdots & \dots \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_{11} & y_{12} & \cdots \\ y_{21} & y_{22} & \cdots \\ \vdots & \vdots & \dots \end{pmatrix}$$

Classical Predictive Models

- Inputs:
  - a set of training samples $\mathcal{D} = \{(\mathrm{x}^i, y^i)\}_{i=1}^N$, where $\mathrm{x}^i = [x_1^i, x_2^i, \cdots, x_d^i]^\top$ and $y^i \in C \triangleq \{c_1, c_2, \cdots, c_L\}$
- Outputs:
  - a predictive function $h(\mathrm{x})$: $y^\star = h(\mathrm{x}) \triangleq \arg\max_y F(\mathrm{x}, y; \mathbf{w})$
- Examples: $F(\mathrm{x}, y; \mathbf{w}) = g(\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y))$

– Logistic Regression, Bayes classifiers
  - Max-likelihood estimation

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

$$p(y|\mathrm{x}) = \frac{\exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y)\}}{\sum_{y'} \exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y')\}}$$

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

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

$$\text{s.t.} \;\; \mathbf{w}^\top \Delta \mathbf{f}_i(y) \geq 1 - \xi_i, \;\; \forall i, \forall y \neq y^i.$$

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 Models

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

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

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp\{\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})\}}{\sum_{\mathbf{y}'} \exp\{\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}')\}}$$

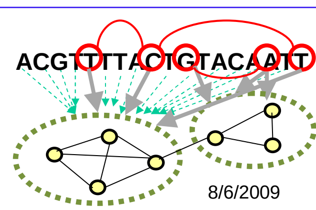
- Max-margin Markov Networks (M^3Ns) (Taskar et al 2003)
  - Based on SVM
  - Max-margin learning (point-estimate)

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

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

where $\mathbf{w}^\top \Delta \mathbf{f}_i(\mathbf{y}|\mathbf{x}_i)$ denotes the margin and $\Delta \ell_i(\mathbf{y})$ is a loss function.

Challenges:
- **SPARSE** prediction model
- Prior information of structures
- Scalable to large-scale problems (e.g., 10^4 input/output dimension)



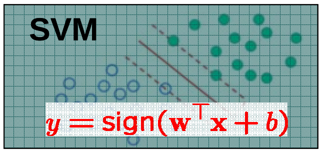
SAILING LAB

Outline

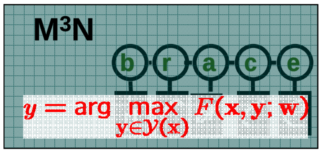
- Structured sparse regression
  - Graph-guided fused lasso: unlinked SNPs to trait networks (Kim and Xing, PLoS Genetics)
  - Temporally-smoothed graph regression: learning time-varying graphs (Ahmed and Xing, PNAS 2009, Kolar and Xing, under review, Annals of Statistics)
- Maximum entropy discrimination Markov networks
  - General Theorems (Zhu and Xing, JMLR submitted)
  - Gaussian MEDN: reduction to M³N (Zhu, Xing and Zhang, ICML 08)
  - Laplace MEDN: a sparse M³N (Zhu, Xing and Zhang, ICML 08)
  - Partially observed MEDN: (Zhu, Xing and Zhang, NIPS 08)
  - Max-margin/Max entropy topic model: (Zhu, Ahmed, and Xing, ICML 09)

SAILING LAB

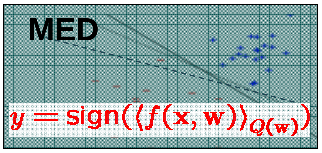
Max-Margin Learning Paradigms



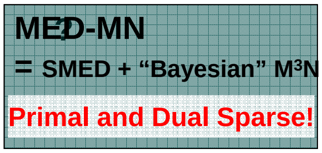
$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m}\xi_i$$
$$y^i(\mathbf{w}^\top \mathbf{x}^i + b) \geq 1 - \xi_i, \quad \forall i$$



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



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



Primal and Dual Problems of M³Ns

- Primal problem:

$$\text{P0 (M}^3\text{N)}: \min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t. } \forall i, \forall \mathbf{y} \neq \mathbf{y}^i: \mathbf{w}^\top \Delta\mathbf{f}_i(\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i,$$
$$\xi_i \geq 0,$$

- Algorithms
  - Cutting plane
  - Sub-gradient
  - …

- Dual problem:

$$\text{D0 (M}^3\text{N)}: \max_{\alpha} \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y}) - \frac{1}{2}\eta^\top\eta$$
$$\text{s.t. } \forall i, \ \forall \mathbf{y}: \sum_{\mathbf{y}} \alpha_i(\mathbf{y}) = C; \ \alpha_i(\mathbf{y}) \geq 0.$$
$$\text{where } \eta = \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y}).$$

- Algorithms:
  - SMO
  - Exponentiated gradient
  - …

$$\mathbf{w}^\star = \eta^\star = \sum_{i,\mathbf{y}} \alpha_i^\star(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y}).$$

- So, M³N is dual sparse!

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 regularization!!

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

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

$$\min \ KL(p(\Theta)\|p_0(\Theta))$$
$$\text{s.t.} \int p(\Theta)[y_i F(x_i;\mathbf{w}) - \xi_i]\,d\Theta \geq 0, \forall i,$$

MED subsumes SVM.

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

SAILING LAB

MaxEnt Discrimination Markov Network

- 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

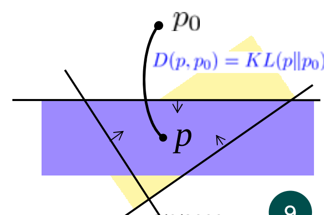
- Feasible subspace of weight distribution:

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

expected margin constraints.

- Average from distribution of M³Ns

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



SAILING LAB

Solution to MaxEnDNet

- Theorem 1:
  - Posterior Distribution:

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

  - Dual Optimization Problem:

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

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

Gaussian MaxEnDNet (reduction to M³N)

- Theorem 2
  - Assume $F(\mathbf{x},\mathbf{y};\mathbf{w}) = \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}), U(\xi) = C\sum_i \xi_i$, and $p_0(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0, I)$
    - Posterior distribution: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}}, I)$, where $\mu_{\mathbf{w}} = \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y})$
    - Dual optimization:
$$\max_{\alpha} \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y}) - \frac{1}{2}\|\sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y})\|^2$$
$$\text{s.t.} \sum_{\mathbf{y}} \alpha_i(\mathbf{y}) = C;\ \alpha_i(\mathbf{y}) \geq 0,\ \forall i,\ \forall \mathbf{y},$$
    - Predictive rule:
$$h_1(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \int p(\mathbf{w})F(\mathbf{x},\mathbf{y};\mathbf{w})\,d\mathbf{w} = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \mu_{\mathbf{w}}^\top \mathbf{f}(\mathbf{x},\mathbf{y})$$
- Thus, MaxEnDNet subsumes M³Ns and admits all the merits of max-margin learning
- Furthermore, MaxEnDNet has at least three advantages …

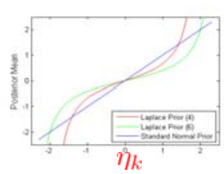
Three Advantages

- An averaging Model: PAC-Bayesian prediction error guarantee

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

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

$$\forall k,\ \langle w_k\rangle_p = \frac{2\eta_k}{\lambda - \eta_k^2}$$



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

I: Generalization Guarantee

- MaxEntNet is an averaging model

- Theorem 3 (PAC-Bayes Bound)

Let p_0 be any continuous probability distribution over $\mathcal{H}$ and $\delta \in (0,1)$

If $\forall F \in \mathcal{H}: \ \mathcal{X} \times \mathcal{Y} \mapsto [-c, c]$

Then with probability at least $1-\delta$ over random samples $\mathcal{D}$ of Q, for very distribution p over $\mathcal{H}$ and for all margin thresholds $\gamma > 0$

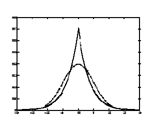
$$\Pr_Q(M(h,\mathbf{x},\mathbf{y}) \le 0) \le \Pr_{\mathcal{D}}(M(h,\mathbf{x},\mathbf{y}) \le \gamma) + O\Big(\sqrt{\frac{\gamma^{-2}KL(p\|p_0)\ln(N|\mathcal{Y}|) + \ln N + \ln \delta^{-1}}{N}}\Big).$$

where $\Pr_Q(.)$ and $\Pr_{\mathcal{D}}(.)$ represent the prob. of event under dist. Q and D, respectively.

II: Laplace MaxEnDNet (primal sparse M^{3}N)

- Laplace Prior:

$$p_0(\mathbf{w}) = \prod_{k=1}^{K} \frac{\sqrt{\lambda}}{2} e^{-\sqrt{\lambda}|w_k|} = \big(\frac{\sqrt{\lambda}}{2}\big)^K e^{-\sqrt{\lambda}\|\mathbf{w}\|}$$

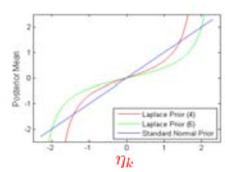


- Corollary 4:
  - Under a Laplace MaxEnDNet, the posterior mean of parameter vector $\mathbf{w}$ is:

$$\forall k, \ \langle w_k \rangle_p = \frac{2\eta_k}{\lambda - \eta_k^2}$$

where the vector η is a linear combination of "support vectors":

$$\eta = \sum_{\alpha} \alpha_i(\mathbf{y}) \Delta \mathbf{f}_i(\mathbf{y})$$



- The Gaussian MaxEnDNet and the regular M^{3}N has no such shrinkage
  - there, we have

$$\langle \mathbf{w} \rangle_p = \eta \iff \forall k, \ \langle w_k \rangle_p = \eta_k$$

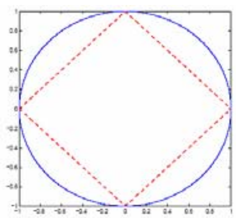
$$\min_{\mu,\xi} \ |\mu| + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t.} \ \ \mu^\top \Delta \mathbf{f}_i(\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i; \ \xi_i \geq 0, \ \ \forall i, \ \forall \mathbf{y} \neq \mathbf{y}^i.$$

LapMEDN vs. L_2 and L_1 regularization

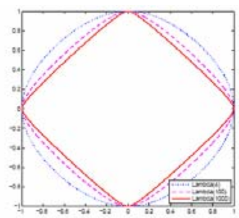
- Corollary 5: LapMEDN corresponding to solving the following primal optimization problem:

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

  - KL norm: $\|\mu\|_{KL} \triangleq \sum_{k=1}^{K}\Big(\sqrt{\mu_k^2+\frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}}\log\frac{\sqrt{\lambda\mu_k^2+1}+1}{2}\Big)$



L_1 and L_2 norms



KL norms

Variational Learning of LapMEDN

- Exact dual function is hard to optimize

$$\max_{\alpha} \ L - \sum_{k=1}^{K}\log\frac{\lambda}{\lambda-\eta_k^2}$$

- Use the hierarchical representation, 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 $\langle\frac{1}{\tau_k}\rangle_q = \sqrt{\frac{\lambda}{\langle w_k^2\rangle_p}}$

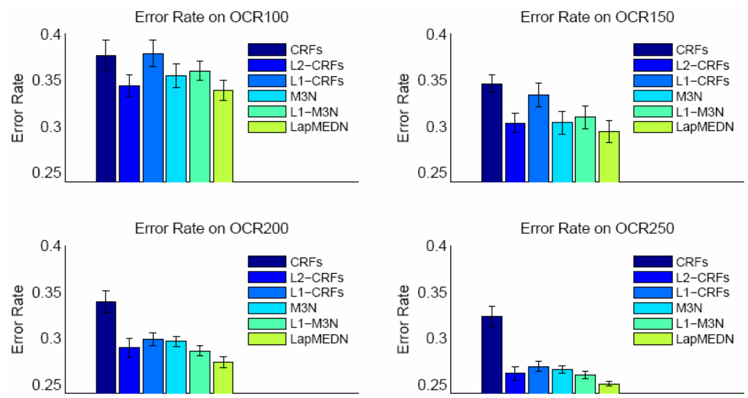
An M^3N optimization problem!

Closed-form solution!

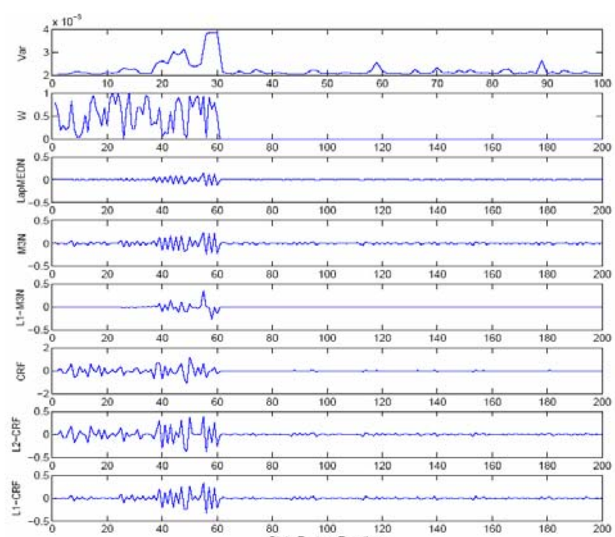
Experimental results on OCR datasets

(CRFs, L_1-CRFs, L_2-CRFs, M^{3}Ns, L_1-M^{3}Ns, and LapMEDN)

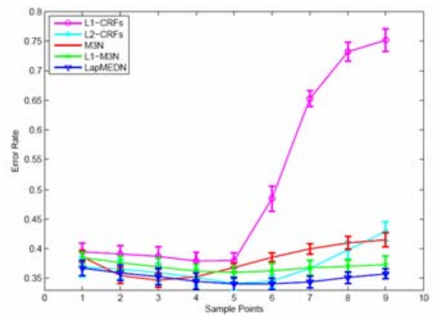
- We randomly construct OCR100, OCR150, OCR200, and OCR250 for 10 fold CV.



Feature Selection



Sensitivity to Regularization Constants



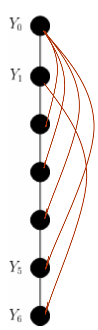
- L_1-CRF and L_2-CRF:
  - 0.001, 0.01, 0.1, 1, 4, 9, 16
- M^{3}N and LapM^{3}N:
  - 1, 4, 9, 16, 25, 36, 49, 64, 81

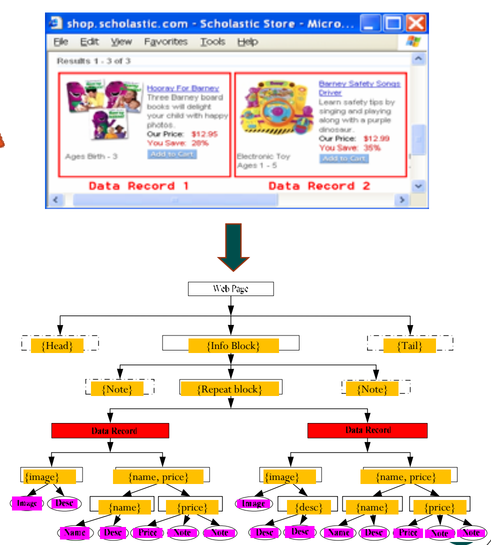
- L_1-CRFs are much sensitive to regularization constants; the others are more stable
- LapM^{3}N is the most stable one

III: Latent Hierarchical MaxEnDNet

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

Given Data Record

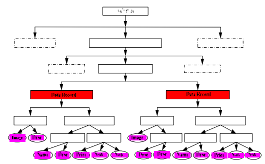




- Hierarchical labeling
- Advantages:
  - Computational efficiency
  - Long-range dependency
  - Joint extraction

Partially Observed MaxEnDNet (PoMEN)

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



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

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

  - 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:

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

- 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
    - M^{3}N problem (QP)
  - o Laplace prior
    - Laplace M^{3}N problem (VB)

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

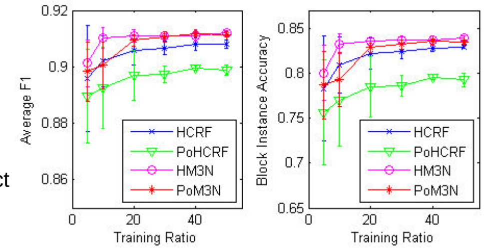
$$\min_{p(\mathbf{w}),\xi} KL(p(\mathbf{z})||p_0(\mathbf{z})) + C\xi_i$$
$$\text{s.t.} \;\; p(\mathbf{z}) \in \mathcal{F}_1^\star, \; \xi_i \geq 0.$$

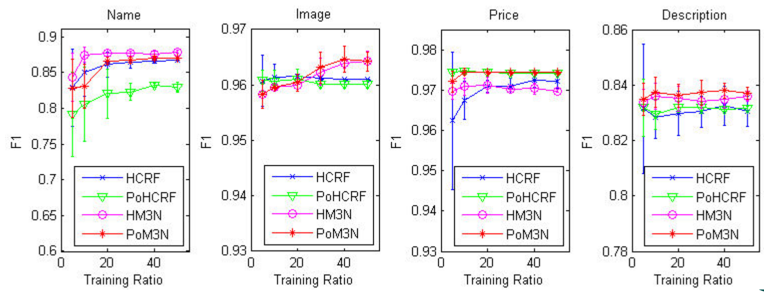
$$\mathcal{F}_1^\star = \{p(\mathbf{z}) : \sum_{\mathbf{z}} p(\mathbf{z}) \int p(\mathbf{w})[\Delta F_i(\mathbf{y},\mathbf{z};\mathbf{w}) - \Delta\ell_i(\mathbf{y})]\,\mathrm{d}\mathbf{w} \geq -\xi_i, \; \forall i, \; \forall \mathbf{y}\}$$

Equivalently reduced to an LP with a polynomial number of constraints

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:





VI: Max-Margin/Max Entropy Topic Model – MED-LDA

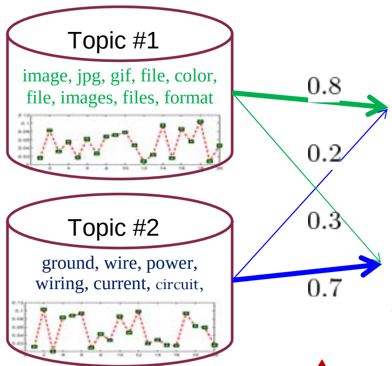
(from images.google.cn)

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

LDA: a generative story for documents

- Bag-of-word representation of documents
- Each word is generated by ONE topic
- Each document is a random mixture over topics



Document #1: gif jpg image current file color images ground power file current format file formats circuit gif images

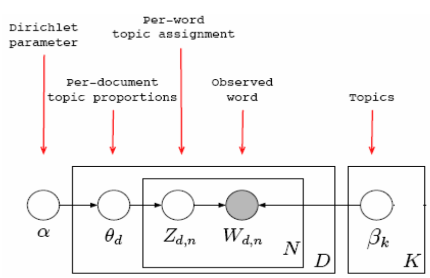
Document #2: wire currents file format ground power image format wire circuit current wiring ground circuit images files…

Mixture Components Mixture Weights

Bayesian Approach → Dirichlet Prior **LDA**

LDA: Latent Dirichlet Allocation

(Blei et al., 2003)



$$\overbrace{\begin{matrix}\theta_1 & \dots & \theta_D \\ 0.8 & \dots & 0.3 \\ 0.2 & \dots & 0.7\end{matrix}}^{\theta} \qquad \beta: \begin{pmatrix} 0.70 & 0.05 & 0.03 & \dots \\ 0.12 & 0.52 & 0.05 & \dots \end{pmatrix}$$

- 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 $W_{d,n} \sim \mathrm{Mult}(\beta_{z_{d,n}})$

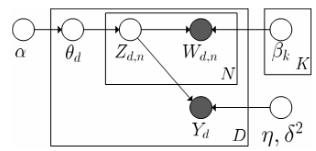
- Joint Distribution: $p(\theta, \mathbf{z}, \mathbf{W}|\alpha, \beta) = \prod_{d=1}^{D} p(\theta_d|\alpha)(\prod_{n=1}^{N} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta))$ exact inference intractable!
- Variational Inference with $q(\mathbf{z}, \theta) \sim p(\mathbf{z}, \theta|\mathbf{W}, \alpha, \beta)$

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

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

Supervised Topic Model (sLDA)

- LDA ignores documents' side information (e.g., categories or rating score), thus lead to suboptimal topic representation for supervised tasks
- Supervised Topic Models handle such problems, e.g., sLDA (Blei & McAuliffe, 2007) and DiscLDA (Simon et al., 2008)
- Generative Procedure (sLDA):
  - 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 $W_{d,n} \sim \mathrm{Mult}(\beta_{z_{d,n}})$
    - Sample y_d



(Blei & McAuliffe, 2007)

$$y_d \sim \begin{cases} \mathcal{N}(\eta^\top \bar{Z}_d, \delta^2), & \text{if } y_d \text{ is continuous} \\ \mathrm{GLM}(\bar{Z}_d, \eta, \delta^2), & \text{otherwise} \end{cases}$$

- Joint distribution:

$$p(\theta, \mathbf{z}, \mathbf{y}, \mathbf{W}|\alpha, \beta, \eta, \delta^2) = \prod_{d=1}^{D} p(\theta_d|\alpha)\Big(\prod_{n=1}^{N} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)\Big)p(y_d|\eta^\top \bar{z}_d, \delta^2)$$

- Variational inference:

$$\mathcal{L}(q) \triangleq -E_q[\log p(\theta, \mathbf{z}, \mathbf{y}, \mathbf{W}|\alpha, \beta, \eta, \delta^2)] - \mathcal{H}(q(\mathbf{z}, \theta)) \geq -\log p(\mathbf{y}, \mathbf{W}|\alpha, \beta, \eta, \delta^2)$$

The big picture

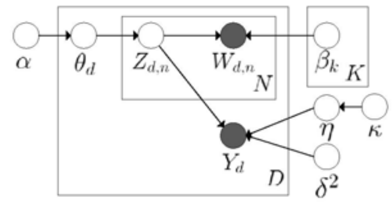
Max-Likelihood Estimation	Max-Margin and Max-Likelihood
sLDA	MedLDA

- How to integrate the max-margin principle into a probabilistic latent variable model?

SAILING LAB

MedLDA Regression Model

- Generative Procedure (Bayesian sLDA):
  - Sample a parameter $\eta \sim p_0(\eta)$
  - 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 $W_{d,n} \sim \mathrm{Mult}(\beta_{z_{d,n}})$
    - Sample y_d: $y_d \sim \mathcal{N}(\eta^\top \bar{Z}_d, \delta^2)$



- Def:

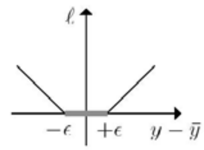
intractable

$$\mathrm{P1(MedLDA}^r) : \min_{\alpha,\beta,\delta^2,\xi,\xi^\star} \boxed{-\log p(\mathbf{y},\mathbf{W}|\alpha,\beta,\delta^2)} + C\sum_{d=1}^{D}(\xi_d+\xi_d^\star)$$

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

model fitting

predictive accuracy



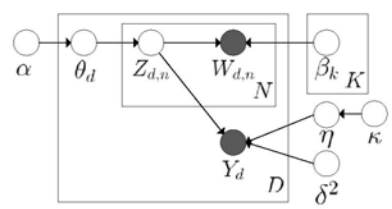
- Predictive Rule:

$$\bar{y} = E[Y|w_{1:N},\alpha,\beta,\delta^2] = E_{q(Z,\eta)}[\eta^\top \bar{Z}|w_{1:N},\alpha,\beta,\delta^2]$$

SAILING LAB

MedLDA Regression Model

- Generative Procedure (Bayesian sLDA):
  - Sample a parameter $\eta \sim p_0(\eta)$
  - 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 $W_{d,n} \sim \mathrm{Mult}(\beta_{z_{d,n}})$
    - Sample y_d: $y_d \sim \mathcal{N}(\eta^\top \bar{Z}_d, \delta^2)$



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

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

  - η is a hidden random variable
  - Variational bound $q(\theta,\mathbf{z},\eta|\gamma,\phi) \sim p(\theta,\mathbf{z},\eta|\alpha,\beta,\delta^2,\mathbf{y},\mathbf{W})$

$$\mathcal{L}(q) \triangleq -E[\log p(\theta,\mathbf{z},\eta,\mathbf{y},\mathbf{W}|\alpha,\beta,\delta^2)] - \mathcal{H}(q(\mathbf{z},\theta,\eta)) \ge -\log p(\mathbf{y},\mathbf{W}|\alpha,\beta,\delta^2)$$

- Predictive Rule:

$$\bar{y} = E[Y|w_{1:N},\alpha,\beta,\delta^2] = E_{q(Z,\eta)}[\eta^\top \bar{Z}|w_{1:N},\alpha,\beta,\delta^2]$$

Variational EM Alg.

- **E-step**: infer the posterior distribution of hidden r.v. $(\theta,\ \mathbf{z},\ \eta)$
- **M-step**: estimate unknown parameters $(\alpha,\ \beta,\ \delta^2)$

- Independence assumption: $q(\theta, \mathbf{z}, \eta | \gamma, \phi) = q(\eta) \prod_{d=1}^{D} q(\theta_d | \gamma_d) \prod_{n=1}^{N} q(z_{dn} | \phi_{dn})$

$$L(\gamma, \phi, q(\eta),\ \alpha, \beta, \delta^2, \xi, \xi^\star, \mu, \mu^\star, v, v^\star) = \mathcal{L}(q) + C \sum_{d=1}^{D} (\xi_d + \xi_d^\star) - \sum_{d=1}^{D} \sum_{i=1}^{N} c_{di} (\sum_{j=1}^{K} \phi_{dij} - 1)$$
$$- \sum_{d=1}^{D} \mu_d (\epsilon + \xi_d - y_d + E[\eta^\top \bar{Z}_d]) - \sum_{d=1}^{D} (\mu_d^\star (\epsilon + \xi_d^\star + y_d - E[\eta^\top \bar{Z}_d]) + v_d \xi_d + v_d^\star \xi_d^\star)$$

- Optimize *L* over ϕ:

$$\phi_{di} \propto \exp \left(\underline{E[\log \theta | \gamma] + E[\log p(w_{di} | \beta)]} + \underline{\frac{y_d}{N\delta^2} E[\eta] - \frac{2E[\eta^\top \phi_{d,-i} \eta] + E[\eta \circ \eta]}{2N^2 \delta^2}} + \underline{\frac{E[\eta]}{N} (\mu_d - \mu_d^\star)} \right)$$

  - The first two terms are the same as in LDA
  - The third and fourth terms are similar to those of sLDA, but in expected version. The variance matters!
  - The last term is a regularizer. Only support vectors affect the topic proportions

- Optimize *L* over other variables. See our paper for details!

MedLDA Classification Model

- Normalization factor in GLM makes inference harder
- We use LDA as the underlying topic model

- Multiclass MedLDA Classification Model:

$$\mathrm{P2(MedLDA^c)}: \min_{q, q(\eta), \alpha, \beta, \xi} \underbrace{\mathcal{L}(q)}_{\text{model fitting}} + \underbrace{KL(q(\eta) || p_0(\eta)) + C \sum_{d=1}^{D} \xi_d}_{\text{predictive accuracy}}$$
$$\text{s.t. } \forall d,\ y \neq y_d:\ E[\eta^\top \Delta \mathbf{f}_d(y)] \geq 1 - \xi_d;\ \xi_d \geq 0,$$

  - Variational upper bound $(q(\theta, \mathbf{z} | \gamma, \phi) \sim p(\theta, \mathbf{z} | \mathbf{W}, \alpha, \beta))$

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

  - *Expected* margin constraints. $\Delta \mathbf{f}_d(y) = \mathbf{f}(y_d, \bar{Z}_d) - \mathbf{f}(y, \bar{Z}_d)$

- Predictive Rule: $y^\star = \arg\max_y E[\eta^\top \mathbf{f}(y, \bar{Z}) | \alpha, \beta]$

Variational EM Alg.

- Independence assumption:

$$q(\theta, \mathbf{z}|\gamma, \phi) = \prod_{d=1}^{D} q(\theta_d|\gamma_d) \prod_{n=1}^{N} q(z_{dn}|\phi_{dn})$$

- Lagrangian function:

$$L(q, q(\eta), \xi, \mu_d(y), c_{di}) = \mathcal{L}(q) + KL(q(\eta)||p_0(\eta)) + C\sum_{d=1}^{D} \xi_d - \sum_{d=1}^{D} v_d \xi_d$$
$$- \sum_{d=1}^{D} \sum_{y \neq y_d} \mu_d(y)(E[\eta^\top \Delta\mathbf{f}_d(y)] + \xi_d - 1) - \sum_{d=1}^{D}\sum_{i=1}^{N} c_{di}(\sum_{j=1}^{K} \phi_{dij} - 1)$$

- Optimize *L* over ϕ:

$$\phi_{di} \propto \exp\Big(\underbrace{E[\log\theta|\gamma] + E[\log p(w_{di}|\beta)]}_{\text{LDA}} + \underbrace{\frac{1}{N}\sum_{y \neq y_d} \mu_d(y) E[\eta_{y_d} - \eta_y]}_{\text{MedLDA}}\Big).$$

Only support vectors matter!

- Optimize *L* over other variables. See the paper for details

MedTM: a general framework

- MedLDA can be generalized to arbitrary topic models:
  - Unsupervised or supervised
  - Generative or undirected random fields (e.g., Harmoniums)

- MED Topic Model (MedTM):

$$\mathrm{P(MedTM)}: \min_{q(H), q(\Upsilon), \Psi, \xi} \underbrace{\mathcal{L}(q(H))}_{\text{model fitting}} + \underbrace{KL(q(\Upsilon)||p_0(\Upsilon)) + U(\xi)}_{\text{predictive accuracy}}$$
$$\text{s.t. } \textit{expected} \text{ margin constraints}$$

- H : hidden r.v.s in the underlying topic model, e.g., $(\theta, \mathbf{z})$ in LDA
- Υ : parameters in predictive model, e.g., η in sLDA
- Ψ : parameters of the topic model, e.g., α in LDA
- $\mathcal{L}$: an variational upper bound of the log-likelihood
- U : a convex function over slack variables

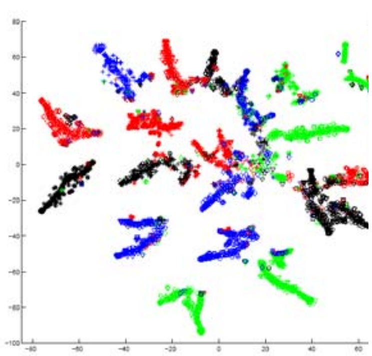
Experiments

- Goal:
  - To qualitatively and quantitatively evaluate how the max-margin estimates of MedLDA affect its topic discovering procedure

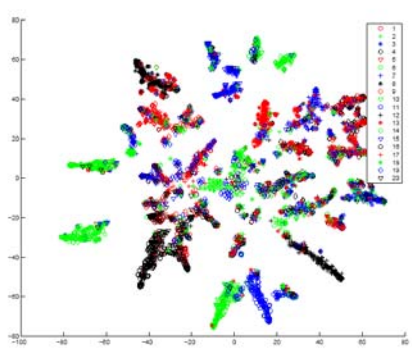
- Data Sets：
  - 20 Newsgroups (classification)
    - Documents from 20 categories
    - ~ 20,000 documents in each group
    - Remove stop word as listed in UMASS Mallet
  - Movie Review (regression)
    - 5006 documents, and 1.6M words
    - Dictionary: 5000 terms selected by tf-idf
    - Preprocessing to make the response approximately normal (Blei & McAuliffe, 2007)

Document Modeling

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



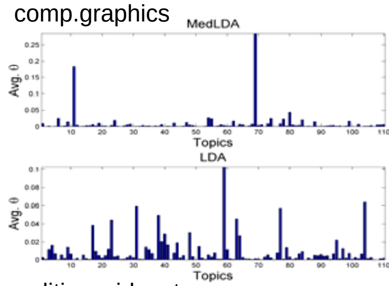
MedLDA



LDA

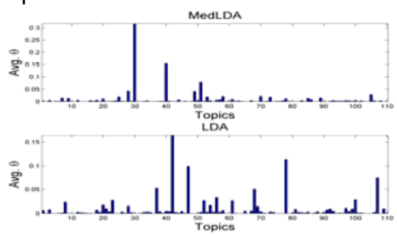
Document Modeling (cont')

comp.graphics



MedLDA			LDA		
T 69	T 11	T 80	T 59	T 104	T 31
image	graphics	db	image	ftp	card
jpeg	image	key	jpeg	pub	monitor
gif	data	chip	color	graphics	dos
file	ftp	encryption	file	mail	video
color	software	clipper	gif	version	apple
files	pub	system	images	tar	windows
bit	mail	government	format	file	drivers
images	package	keys	bit	information	vga
format	fax	law	files	send	cards
program	images	escrow	display	server	graphics

politics.mideast

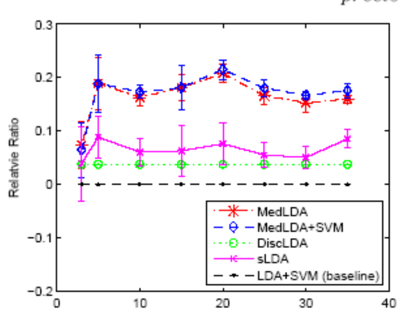


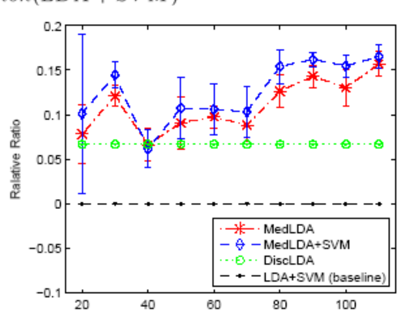
T 30	T 40	T 51	T 42	T 78	T 47
israel	turkish	israel	israel	jews	armenian
israeli	armenian	lebanese	israeli	jewish	turkish
jews	armenians	israeli	peace	israel	armenians
arab	armenia	lebanon	writes	israeli	armenia
writes	people	people	article	arab	turks
people	turks	attacks	arab	people	genocide
article	greek	soldiers	war	arabs	russian
jewish	turkey	villages	lebanese	center	soviet
state	government	peace	lebanon	jew	people
rights	soviet	writes	people	nazi	muslim

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

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

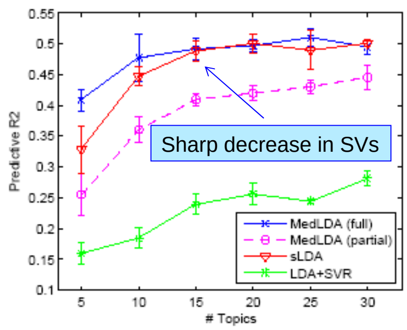


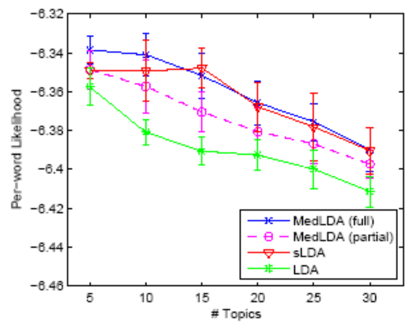


Regression

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

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

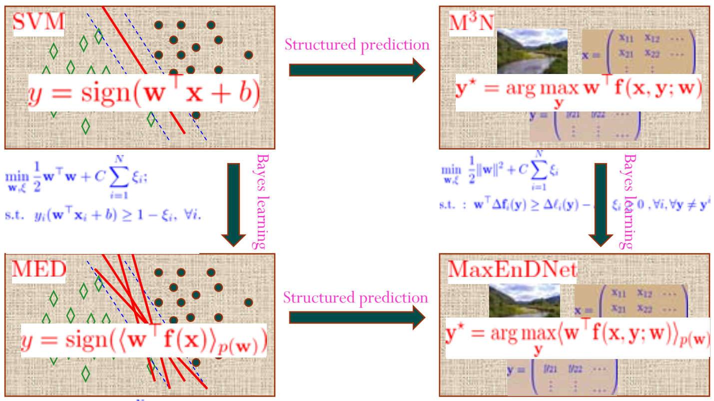




Summary

- 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
- Laplace MaxEnDNet: simultaneously primal and dual sparse
  - Can perform as well as sparse models on synthetic data
  - Perform better on real data sets
  - More stable to regularization constants
- PoMEN
  - Provides an elegant approach to incorporate latent variables and structures under max-margin framework
  - Experimental results show the advantages of max-margin learning over likelihood methods with latent variables

Margin-based Learning Paradigms



$$\min_{\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{i=1}^{N}\xi_i;$$
$$\text{s.t. } y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 - \xi_i, \ \forall i.$$

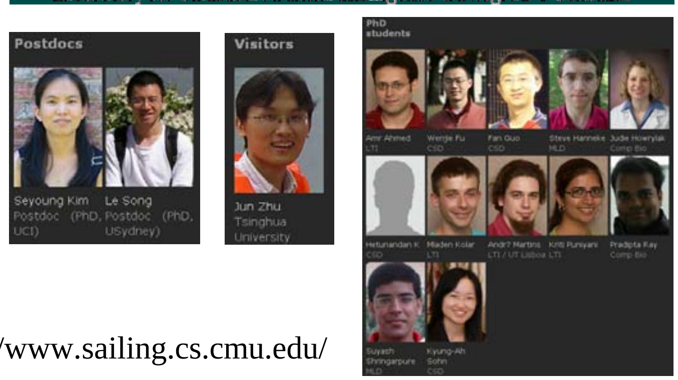
$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t. : } \mathbf{w}^\top\Delta\mathbf{f}_i(\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i, \xi_i \geq 0, \forall i, \forall \mathbf{y} \neq \mathbf{y}^i$$

$$\min_{p,\xi} KL(p\|p_0) + C\sum_{i=1}^{N}\xi_i;$$
$$\text{s.t. } y_i\langle\mathbf{f}(\mathbf{x}_i)\rangle_{p(\mathbf{w})} \geq 1 - \xi_i, \ \forall i.$$

$$\min_{p(\mathbf{w}),\xi} KL(p\|p_0) + U(\xi)$$
$$\text{s.t. } \int p(\mathbf{w})[\Delta F_i(\mathbf{y};\mathbf{w}) - \Delta\ell_i(\mathbf{y})]\,d\mathbf{w} \geq -\xi_i, \ \xi_i \geq 0, \ \forall i, \forall \mathbf{y} \neq \mathbf{y}^i.$$

Acknowledgement

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics



http://www.sailing.cs.cmu.edu/

Thanks!

Reference:

Markov Chain Prior

$$P(\mathbf{c}) = P(c_1) \prod_{j=2}^{J} P(c_j | c_{j-1})$$

Markov Chain Prior

$$P(\mathbf{c}) = P(c_1) \prod_{j=2}^{J} \underline{P(c_j | c_{j-1})}$$

- $c_j = c_{j-1}$ if
  1) the distance between the two SNPs is small, or
  2) the recombination rate between the two SNPs is small

Markov Chain Prior

$$P(\mathbf{c}) = P(c_1) \prod_{j=2}^{J} \underline{P(c_j | c_{j-1})}$$

Poisson process

$$P(c_j | c_{j-1}) = \exp(-d_j \rho_j)\ \delta(c_j, c_{j-1}) + (1 - \exp(-d_j \rho_j))\ \Pi_{c_{j-1}, c_j}$$

- ρ_j : Recombination rate at *j*th SNP
- d_j : Distance between *j*th and (*j*-1)th SNP
- Π : Transition probability matrix

$$\begin{pmatrix} \pi_0 & 1 - \pi_0 \\ 1 - \pi_1 & \pi_1 \end{pmatrix}$$

Variational Bayesian Learning (Cont')

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

Initialize $\langle \mathbf{w}\rangle_p^1 \leftarrow 0, \Sigma_{\mathbf{w}}^1 \leftarrow I$

$$p(\mathbf{w}) \propto \exp\{\int q(\tau)\log p(\mathbf{w}|\tau)\,\mathrm{d}\tau - b\}\cdot\exp\{\mathbf{w}^\top\eta - L\}$$
$$\propto \exp\{-\frac{1}{2}\mathbf{w}^\top\langle A^{-1}\rangle_q\mathbf{w} - b + \mathbf{w}^\top\eta - L\}$$
$$= \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}},\Sigma_{\mathbf{w}}),\ \text{where } L = \sum_{i,\mathbf{y}}\alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y}),\ A = \mathrm{diag}(\tau_k).$$

Solve an M^3N Problem ($\Sigma_{\mathbf{w}}^t$)

$$\mu_{\mathbf{w}} = \Sigma_{\mathbf{w}}\eta \text{ and } \Sigma_{\mathbf{w}} = (\boxed{\langle A^{-1}\rangle_q})^{-1} = \langle\mathbf{w}\mathbf{w}^\top\rangle_p - \boxed{\langle\mathbf{w}\rangle_p\langle\mathbf{w}\rangle_p^\top}$$

$$\max_\alpha \sum_{i,\mathbf{y}}\alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y}) - \frac{1}{2}\eta^\top\Sigma_{\mathbf{w}}\eta$$
$$\text{s.t. } \sum_{\mathbf{y}}\alpha_i(\mathbf{y}) = C;\ \alpha_i(\mathbf{y}) \geq 0,\ \forall i,\ \forall \mathbf{y}.$$

$$\min_{\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^\top\Sigma_{\mathbf{w}}^{-1}\mathbf{w} + C\sum_{i=1}^N \xi_i$$
$$\text{s.t. } \mathbf{w}^\top\Delta\mathbf{f}_i(\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i;\ \xi_i \geq 0,\ \forall i,\ \forall\mathbf{y}\neq\mathbf{y}^i.$$

$t \leftarrow t+1$

Update $\Sigma_{\mathbf{w}}^{t+1}$

$$q(\tau) = \prod_{k=1}^K q(\tau_k)$$

$$q(\tau_k) \propto p(\tau_k|\lambda)\exp\{\langle\log p(w_k|\tau_k)\rangle_p\}$$
$$\propto \mathcal{N}(\sqrt{\langle w_k^2\rangle_p}|0,\tau_k)\exp(-\frac{1}{2}\lambda\tau_k).$$

$$\boxed{\langle\frac{1}{\tau_k}\rangle_q} = \int\frac{1}{\tau_k}q(\tau_k)\,\mathrm{d}\tau_k = \sqrt{\frac{\lambda}{\boxed{\langle w_k^2\rangle_p}}}$$

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