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
\[ x = (x_{11}, x_{12}, \ldots) \quad y = y_i \]

- Structured prediction
  - Part of speech tagging
  \[ x = \text{"Do you want sugar in it?"} \quad y = \text{<verb pron verb noun prep pron>} \]
  - Image segmentation
\[ x = \begin{pmatrix} x_{11} & x_{12} & \ldots \\ x_{21} & x_{22} & \ldots \\ \vdots & \vdots & \vdots \end{pmatrix} \quad y = \begin{pmatrix} y_{11} & y_{12} & \ldots \\ y_{21} & y_{22} & \ldots \\ \vdots & \vdots & \vdots \end{pmatrix} \]
Laplace Max-margin Markov Networks

Classical Predictive Models

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

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

- Logistic Regression, Bayes classifiers
  - Max-likelihood estimation
  
  \[ \max_{\mathbf{w}} \mathcal{L}(\mathcal{D}; \mathbf{w}) = \sum_{i=1}^N \log p(y^i|x^i) \]
  \[ p(y|x) = \frac{\exp\{\mathbf{w}^T f(x, y)\}}{\sum_{y'} \exp\{\mathbf{w}^T f(x, y')\}} \]

- Support Vector Machines (SVM)
  - Max-margin learning
  
  \[ \min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i; \]
  \[ \text{s.t. } \mathbf{w}^T \Delta f_i(y) \geq 1 - \xi_i, \forall i, \forall y \neq y^i. \]

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)
  
  \[ \max_{\mathbf{w}} \mathcal{L}(\mathcal{D}; \mathbf{w}) = \sum_{i=1}^N \log p(y|x^i) \]
  \[ p(y|x) = \frac{\exp\{\mathbf{w}^T f(x, y)\}}{\sum_{y'} \exp\{\mathbf{w}^T f(x, y')\}} \]

- Max-margin Markov Networks (M^3Ns) (Taskar et al 2003)
  - Based on SVM
  - Max-margin learning (point-estimate)
  
  \[ \min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i; \]
  \[ \text{s.t. } \forall i, \forall y \neq y^i : \mathbf{w}^T \Delta f_i(y) \geq \Delta f_i(y^i) - \xi_i, \xi_i \geq 0, \]

where \( \mathbf{w}^T \Delta f_i(y|x_i) \) denotes the margin and \( \Delta f_i(y) \) is a loss function.

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

ACGTTTTACTGTACAATT

VLPR 2009 @ Beijing, China
Outline

- Structured sparse regression
  - Graph-guided fused lasso: unlinked SNPs to trait networks (Kim and Xing, PLoS Genetics)

- Maximum entropy discrimination Markov networks
  - General Theorems (Zhu and Xing, JMLR submitted)
  - Gaussian MEDN: reduction to M^3N (Zhu, Xing and Zhang, ICML 08)
  - Laplace MEDN: a sparse M^3N (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)

Max-Margin Learning Paradigms
Primal and Dual Problems of $M^3Ns$

- **Primal problem:**
  \[ P_0 (M^3N) : \min_w \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \]
  such that, \( \forall i, \forall y \neq y^i : w^\top \Delta f_i(y) \geq \Delta f_i(y)^i - \xi_i \)
  \( \xi_i \geq 0 \)

- **Algorithms**
  - Cutting plane
  - Sub-gradient
  - ...

- **Dual problem:**
  \[ D_0 (M^3N) : \max_\alpha \sum_{i,y} \alpha_i(y) \Delta f_i(y) - \frac{1}{2} y^\top \eta \]
  such that, \( \forall i, \forall y : \sum_{i,y} \alpha_i(y) = C; \ \alpha_i(y) \geq 0 \)
  where \( \eta = \sum_{i,y} \alpha_i(y) \Delta f_i(y) \)

- **Algorithms:**
  - SMO
  - Exponentiated gradient
  - ...

\[ w^* = \eta^* = \sum_{i,y} \alpha^*_i(y) \Delta f_i(y). \]

- So, $M^3N$ 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
  - The optimization problem (binary classification)
  \[ \min_{\theta} \frac{1}{n} \sum_{i=1}^n KL(p(\theta) || p_i(\theta)) \]
  \[ \text{MED subsumes SVM.} \]

\[ \hat{g} = \text{sign} \int p(w) F(x; w) dw \quad (y \in \{+1, -1\}) \]

where $\Theta$ is the parameter $w$ when $\xi$ are kept fixed or the pair $(w, \xi)$ when we want to optimize over $\xi$.
MaxEnt Discrimination Markov Network

- **Structured MaxEnt Discrimination (SMED):**

\[
P_1 : \min_{p(w), \xi} \text{KL}(p(w) || p_0(w)) + U(\xi)
\]

\[\text{s.t. } p(w) \in \mathcal{F}_1, \xi_i \geq 0, \forall i.\]

\text{generalized maximum entropy or regularized KL-divergence}

- **Feasible subspace of weight distribution:**

\[
\mathcal{F}_1 = \{p(w) : \int p(w) (\Delta F_i(y; w) - \Delta F_i(y')) dw \geq -\xi_i, \forall y \neq y', \text{ expected margin constraints.}\}
\]

- **Average from distribution of M3Ns**

\[
h_1(x; p(w)) = \arg \max_{y \in \mathcal{Y}(x)} \int p(w) F(x, y; w) dw
\]

---

Solution to MaxEnDNet

- **Theorem 1:**

  - Posterior Distribution:

\[
p(w) = \frac{1}{Z(\alpha)} p_0(w) \exp \left\{ \sum_{i,y} \alpha_i(y) [\Delta E_i(y; w) - \Delta E_i(y)] \right\}
\]

  - Dual Optimization Problem:

\[
D_1 : \max_{\alpha} - \log Z(\alpha) - U^*(\alpha)
\]

\[\text{s.t. } \alpha_i(y) > 0, \forall y, \forall \alpha.\]

\[U^*(\cdot) \text{ is the conjugate of the } U(\cdot), \text{ i.e., } U^*(\alpha) = \sup_{\xi} \left( \sum_{i,y} \alpha_i(y) \xi_i - U(\xi) \right)\]
Gaussian MaxEnDNet (reduction to M\(^3\)N)

- **Theorem 2**
  - Assume \( F(x, y; w) = \mathbf{w}^T \Phi(x, y), U(\xi) = C \sum \xi_i \) and \( p_0(w) = \mathcal{N}(w|0, I) \)
  - Posterior distribution: \( p(w) = \mathcal{N}(w|\mu_w, \Sigma) \)
  - Dual optimization:
    \[
    \max_{\alpha} \sum_{y \in Y} \alpha(y) \Delta w(y) - \frac{1}{2} \sum_{y \in Y} \alpha(y) \Delta L(y)|\|^2 \\
    \text{s.t.} \sum_{y \in Y} \alpha(y) = C, \alpha(y) \geq 0, \forall y
    \]
  - Predictive rule:
    \[
    h_\ell(x) = \arg \max_{y \in Y(x)} \int p(w) F(x, y; w) dw = \arg \max_{y \in Y(x)} \mu_w \Phi(x, y)
    \]

- Thus, MaxEnDNet subsumes M\(^3\)Ns 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
  \[
  Pr_Q(M(h(x, y)) \leq \epsilon) \leq Pr_P(M(h(x, y)) \leq \epsilon) + O\left( \frac{K e^{-2KL(p_{\|\alpha}\|N(0, I))} + \ln N + K \delta^{-1}}{N} \right).
  \]
- Entropy regularization: Introducing useful biases
  - Standard Normal prior \( \Rightarrow \) reduction to standard M\(^3\)N (we’ve seen it)
  - Laplace prior \( \Rightarrow \) Posterior shrinkage effects (sparse M\(^3\)N)
    \[
    \forall k, \langle \eta_k \rangle_p = \frac{2\nu_k}{\lambda - \eta_k^2}
    \]
- 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 $\delta \in (0, 1]$. Let $\forall F \in \mathcal{H} : X \times Y \mapsto [-c, c]$

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

  \[
  \Pr_Q(M(h, x, y) \leq 0) \leq \Pr_p(M(h, x, y) \leq \gamma) + O\left(\frac{\gamma^2 KL(p || p_0) \ln(N) \gamma}{N} + \ln N + \ln \frac{1}{\delta}\right),
  \]

  where $\Pr_Q(\cdot)$ and $\Pr_p(\cdot)$ represent the probability of event under dist. $Q$ and $D$, respectively.

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

- Laplace Prior:

  \[
  p_0(w) = \prod_{k=1}^{K} \frac{\sqrt{\lambda}}{2} e^{-\sqrt{\lambda} |w_k|} = \left(\frac{\sqrt{\lambda}}{2}\right)^K e^{-\sqrt{\lambda} \|w\|_1}
  \]

- Corollary 4:

  - Under a Laplace MaxEnDNet, the posterior mean of parameter vector $w$ is:

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

    where the vector $\eta$ is a linear combination of "support vectors":

    \[
    \eta = \sum_{i} \alpha_i(y) \Delta_i(y)
    \]

  - The Gaussian MaxEnDNet and the regular $M^3N$ has no such shrinkage

    - there, we have

    \[
    \langle w \rangle_p = \eta \iff \forall k, \quad \langle w_k \rangle_p = \eta_k
    \]
Laplace Max-margin Markov Networks

LapMEDN vs. L₂ and L₁ regularization

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

\[
\min_{\mu, \xi} \sqrt{\lambda} \sum_{k=1}^{K} \left( \sqrt{p_k^2 + \frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}} \log \frac{\sqrt{\lambda} p_k^2 + 1 + 1}{2} \right) + C \sum_{i=1}^{N} \xi_i
\]

s.t. \( \mu^T \Delta f_i(y) \geq \Delta f_i(y) - \xi_i; \ \xi_i \geq 0, \ \forall i, \ \forall y \neq y_i \).

- KL norm: \( \|p\|_{KL} = \sum_{i=1}^{K} \left( \sqrt{p_k^2 + \frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}} \log \frac{\sqrt{\lambda} p_k^2 + 1 + 1}{2} \right) \)

Variational Learning of LapMEDN

- Exact dual function is hard to optimize

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

- Use the hierarchical representation, we get:

\[
KL(p||p_0) = -H(p) - (\log \int p(w|\tau)p(\tau|\lambda) d\tau)_p
\]

\[
\leq -H(p) - (\int q(\tau) \log \frac{p(w|\tau)p(\tau|\lambda)}{q(\tau)} d\tau)_p \triangleq \mathcal{L}(p(w), q(\tau))
\]

- We optimize an upper bound:

\[
\min_{p(w) \in \mathcal{F}(q(\tau))} \mathcal{L}(p(w), q(\tau)) + U(\xi)
\]

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

Keep \( q(\tau) \) fixed

Keep \( p(w) \) fixed

The effective prior is normal

Closed form solution of \( q(\tau) \) and its expectation

\[
\forall k: \ p_0(w_k | \tau_k) = \mathcal{N}(w_k | \mathcal{N}(0, 1), \frac{1}{\tau_k} q(\tau) )
\]

\[
\mathbb{E}[q(\tau)|p] = \frac{1}{\tau_k} \mathbb{E}[\sqrt{\frac{1}{\lambda} \log \frac{\sqrt{\lambda} p_k^2 + 1 + 1}{2}} | p]
\]
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

- $L_1$-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.

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

Given Data Record
Partially Observed MaxEnDNet (PoMEN)

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

  - PoMEN: learning \( p(w, z) \)

  \[
  \text{P2(PoMEN)}: \min_{p(w,z) \in \mathcal{F}_2} \frac{1}{\xi_i} KL(p(w,z)||p_0(w(z))) + U(\xi)
  \]

  \[
  \mathcal{F}_2 = \{ p(w,z) : \sum \int p(w,z) |\Delta F_i(y;x,w) - \Delta F_i(y) || dw \geq -\xi_i, \forall i, \forall y \neq y' \}.
  \]

  - Prediction:

    \[
    h_2(x) = \arg \max_{y \in \Delta(x)} \sum \int p(w,z) F(x,y,z;w) \, dw
    \]

---

Alternating Minimization Alg.

- Factorization assumption:

  \[
  p_0(w,z) = p_0(w) \prod_{i=1}^{N} p_0(z_i), \quad p(w,z) = p(w) \prod_{i=1}^{N} p(z_i)
  \]

- Alternating minimization:

  - Step 1: keep \( p(z) \) fixed, optimize over \( p(w) \)

    \[
    \min_{p(w) \in \mathcal{F}'} KL(p(w)||p_0(w)) + C \sum \xi_i
    \]

    \[
    \mathcal{F}' = \{ p(w) : \int p(w) F_i(y;x,w) - \Delta F_i(y) || dw \geq -\xi_i, \forall y \}
    \]

  - Step 2: keep \( p(w) \) fixed, optimize over \( p(z) \)

    \[
    \min_{p(z) \in \mathcal{F}''} KL(p(z)||p_0(z)) + C \xi_i
    \]

    \[
    \mathcal{F}'' = \{ p(z) : \sum p(z) \int p(w) |\Delta F_i(y;x,w) - \Delta F_i(y) || dw \geq -\xi_i, \forall y \}
    \]

  \( \mathcal{F} = \mathcal{F}' \cap \mathcal{F}'' \)

  - Normal prior
  - Laplace prior

  \( M^N \) problem (QP)

  \( M^N \) problem (VB)

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)
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

```
Topic #1
image, jpg, gif, file, color, file, images, files, format

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

0.8

0.2

Topic #2
ground, wire, power, wiring, current, circuit,

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

0.3

0.7
```

LDA: Latent Dirichlet Allocation

(Blei et al., 2003)

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

- Joint Distribution:
  \[
  p(\theta, z, W | \alpha, \beta) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} p(\theta_d | \alpha) p(z_{d,n} | \theta_d) p(W_{d,n} | z_{d,n}, \beta)
  \]

- Variational Inference with $q(z, \theta) \sim (z, \theta|\alpha, \beta)$
  \[
  \mathcal{L}(q) \triangleq -E_q[\log p(\theta, z, W | \alpha, \beta)] - \mathcal{H}(q(z, \theta)) > -\log p(W | \alpha, \beta)
  \]

- 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 $\theta_d \sim \text{Dir}(\alpha)$
    - For each word:
      - Sample a topic $Z_{d,n} \sim \text{Mult}(\theta_d)$
      - Sample a word $W_{d,n} \sim \text{Mult}(\beta_{Z_{d,n}})$
  - Sample $y_d$

Generative Procedure (sLDA):

- Joint distribution:
  \[ p(\theta, z, y, W|\alpha, \beta, \eta, \delta^2) = \prod_d p(\theta_d|\alpha) \prod_{i=1}^N \prod_n p(z_{d,n}|\theta_d)p(w_{d,n}|z_{d,n}, \theta)p(y_d|\eta^Tz_d, \delta^2) \]

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

The big picture

<table>
<thead>
<tr>
<th>Max-Likelihood Estimation</th>
<th>Max-Margin and Max-Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>sLDA</td>
<td>MedLDA</td>
</tr>
</tbody>
</table>

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

- **Generative Procedure (Bayesian sLDA):**
  - Sample a parameter \( \eta \sim p(\eta) \)
  - For each document \( d \):
    - Sample a topic proportion \( \theta_d \sim \text{Dir}(\alpha) \)
    - For each word:
      - Sample a word \( Z_d, n \sim \text{Mult}(\theta_d) \)
      - Sample \( \omega_d \sim N(\eta^T Z_d, \delta^2) \)
  - Sample \( \mu_d \sim \mathcal{N}(\eta^T Z_d, \delta^2) \)

- **Def:**
  \[
P_1(\text{MedLDA}^+): \quad \min_{\alpha, \beta, \delta^2, \xi^d} \quad -\log p(y, W|\alpha, \beta, \delta^2) \leq \varepsilon + \xi_d, \mu_d \]
  \[
  \text{s.t. } \forall d : \quad \xi_d \geq 0, \mu_d \geq 0, v_d \geq 0 \]

- **Predictive Rule:**
  \[
  \hat{y} = E[Y|w_{1:N}, \alpha, \beta, \delta^2] = E_{q(Z, \eta)}[\eta^T Z|w_{1:N}, \alpha, \beta, \delta^2] \]

\[8/6/2009\] VLPR 2009 @ Beijing, China
Variational EM Alg.

- **E-step**: infer the posterior distribution of hidden r.v. \( y, z, \eta \)
- **M-step**: estimate unknown parameters \( (\alpha, \beta, \delta) \)

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

\[
L(\gamma, \alpha, \phi, \beta, \delta, \xi, \mu, \nu, \epsilon \gamma^* = \xi^* \mu + \nu \xi^* + \epsilon) = \mathcal{L}(\eta) + C \sum_{d=1}^{D} (\zeta_d - 1) - \sum_{n=1}^{N} \sum_{j=1}^{K} \phi_{dnj} - \sum_{d=1}^{D} \xi_d (\mu_d - 1) \]

- Optimize \( L \) over \( \phi_{ij} \):
  \[
  \phi_{ij} \propto \exp \left( \mathbb{E}[\log \theta_{ij}] + \mathbb{E}[\log p(n_{ij}|\beta)] \right) + \frac{\mu_d \xi_d E[n]}{N} - \frac{2 \mathbb{E}[\eta^T \delta_u - \eta] + \mathbb{E}[\eta \eta^T]}{2N^2 \delta_u^2} + \frac{\mathbb{E}[\eta^T \eta]}{N (\mu^T - \mu^*)}
  \]
  - 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:
  \[
  P(\text{MedLDA}_c) : \min_{\eta, \phi, \gamma, \alpha, \beta, \delta} \mathcal{L}(\eta) + KL(q(\eta) || p_\theta(\eta)) + C \sum_{d=1}^{D} \xi_d \\
  \text{s.t. all } y \neq y_d : E[\eta^T \pi_d(y)] > 1 - \xi_d; \xi_d > 0
  \]

  - Variational upper bound \( (q(\theta, z, \eta, \phi) \sim p(\theta, z|\mathbf{W}, \alpha, \beta)) \)
  \[
  \mathcal{L}(\eta) \overset{\text{var}}{=} -E[\log p(\theta, z|\mathbf{W}|\alpha, \beta)] - H(q(\theta, z)) \geq -\log p(\mathbf{W}|\alpha, \beta)
  \]

  - **Expected** margin constraints. \( \Delta \theta_d(y) = \pi(y_d, \tilde{Z}_d) - \pi(y, \tilde{Z}_d) \)

- **Predictive Rule**:
  \[
  y^* = \arg \max_y E[\eta^T \pi(y, \tilde{Z})|\alpha, \beta]
  \]
Variational EM Alg.

- Independence assumption:
  \[ q(\theta, x|\gamma, \phi) = \prod_{j=1}^{D} q(\theta_d|\gamma_d) \prod_{n=1}^{N} q(z_{dn}|\phi_{dn}) \]

- Lagrangian function:
  \[
  L(q, q(\eta), \mu_d(y), \phi_d) = \mathcal{L}(q) + KL(q(\eta)||p_0(\eta)) + C \sum_{d=1}^{D} \xi_d - \sum_{d=1}^{D} v_d \xi_d \\
  + \sum_{d=1}^{D} \sum_{y \neq y_{ld}} \mu_d(y)(E[\eta^T \Delta f_d(y)] + \xi_d - 1) - \sum_{d=1}^{D} \sum_{i=1}^{N} c_{di}(\sum_{j=1}^{K} \phi_{di,j} - 1)
  \]

- Optimize \( L \) over \( \phi_d \):
  \[
  \phi_{di} \propto \exp(E[\log \theta] + E[\log p(w_{di} | \beta)] + \frac{1}{N} \sum_{y} \mu_d(y)E[\eta_{ld} - \eta_y]).
  \]

- 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):
  \[
  P(\text{MedTM}) : \min_{q(H), q(\gamma), \xi} \mathcal{L}(q(H)) + KL(q(\gamma)||p_0(\gamma)) + U(\xi) \\
  \text{s.t. expected margin constraints}
  \]

  - \( H \): hidden r.v.s in the underlying topic model, e.g., \((\theta, z)\) in LDA
  - \( \gamma \): parameters in predictive model, e.g., \( \gamma \) in sLDA
  - \( \xi \): parameters of the topic model, e.g., \( \xi \) 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** (van der Maaten & Hinton, 2008)
**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)), MedLDA, MedLDA+SVM
- **Measure**: Relative Improvement Ratio
  
  \[
  RR(M) = \frac{\text{precision}(M)}{\text{precision}(LDA + SVM)} - 1
  \]
Regression

- **Data Set**: Movie Review (Blei & McAuliffe, 2007)
- **Models**: MedLDA\textit{(partial)}, MedLDA\textit{(full)}, sLDA, LDA+SVR
- **Measure**: predictive R\textsuperscript{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\textsuperscript{3}Ns
  - Model averaging: PAC-Bayes theoretical error bound
  - Entropic regularization: sparse M\textsuperscript{3}Ns
  - 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

- SVM
  \[ y = \text{sign}(w^T x + b) \]
  \[ \min \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i \]
  \[ \text{s.t. } \xi_i(w^T x_i + b) \geq 1 - \xi_i, \forall i \]

- MRF
  \[ y = \text{sign}(\sum_{w} f(x)_w p(w)) \]
  \[ \min_{p(w)} K(p|\pi) + C \sum_{w} \xi_w \]
  \[ \text{s.t. } \pi(w)p(w) \geq 1 - \xi_w, \forall w \]

- Laplace Max-margin Markov Networks
  \[ y^* = \arg \max_{y} \sum_{y'} \max_{w} f(x, y, y')_w \]
  \[ \min_{w} \frac{1}{2} ||w||^2 + C \sum_{w} \xi_w \]
  \[ \text{s.t. } w^T M(y) \geq \Delta(y), \forall y, y \neq y' \]

Acknowledgement

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

Reference:

Markov Chain Prior

\[ P(c) = P(c_1) \prod_{j=2}^{J} P(c_j | c_{j-1}) \]
Markov Chain Prior

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

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

Markov Chain Prior

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

Poisson process

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

- \( \rho_j \): Recombination rate at \( j \)th SNP
- \( d_j \): Distance between \( j \)th and \((j-1)\)th SNP
- \( \Pi \): Transition probability matrix

\[
\begin{pmatrix}
\pi_0 & 1 - \pi_0 \\
1 - \pi_1 & \pi_1
\end{pmatrix}
\]
Variational Bayesian Learning (Cont')

\[ \min_{p(w) \in P_1, q(\tau) \in \xi} L(p(w), q(\tau)) + U(\xi) \]

Initialize \( \langle w \rangle_0 \leftarrow 0, \Sigma_w^1 \leftarrow I \)

Solve an \( \Sigma^M \)N Problem \( \Sigma^M_w \)

\[ t \leftarrow t + 1 \]

Update \( \Sigma^2_w \)

Variational Bayesian Learning (Cont')

Experimental Results

- Web data extraction:
  - Name, Image, Price, Description

- Methods:
  - Hierarchical CRFs, Hierarchical \( M^3 \)N
  - 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

VLPR 2009 @ Beijing, China
8/6/2009