Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

On Learning Sparse Structured Input-Output Models

Eric Xing

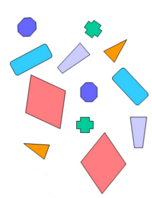
epxing@cs.cmu.edu

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

Carnegie Mellon University

Acknowledgement:
Amr Ahmed, Jacob Eisenstein, Xi Chen, Qirong Ho, Seyoung Kim, Seunghak Lee, Andre Martins, Noah Smith, and Jun Zhu

Unstructured Prediction Problem



$$\mathbf{x} = \left(\mathrm{x}_{11} \quad \mathrm{x}_{12} \quad \ldots \right) \quad \Rightarrow \quad \mathbf{y} = y_1$$

SAILING LAB

Classical Predictive Models

- Input and output space: $\mathcal{X} \triangleq \mathbb{R}^{M_x}$ $\quad \mathcal{Y} \triangleq \{-1,+1\}$
- Predictive function $h(\mathrm{x})$: $y^\star = h(\mathrm{x}) \triangleq \arg\max_{y \in \mathcal{Y}} F(\mathrm{x}, y; \mathbf{w})$
- Examples: $F(\mathrm{x}, y; \mathbf{w}) = g(\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y))$
- Learning: $\hat{\mathrm{w}} = \arg\min_{\mathbf{w} \in \mathcal{W}} \ell(\mathrm{x}, y; \mathbf{w}) + \lambda R(\mathbf{w})$

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

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

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

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

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

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

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

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

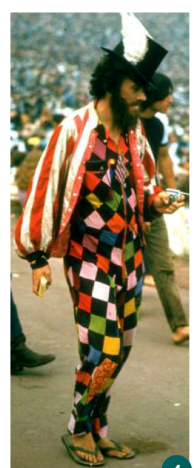
SAILING LAB

From Unstructured to Structured Prediction

- Binary classification: black-and-white decisions

- Multi-class classification: the world of technicolor

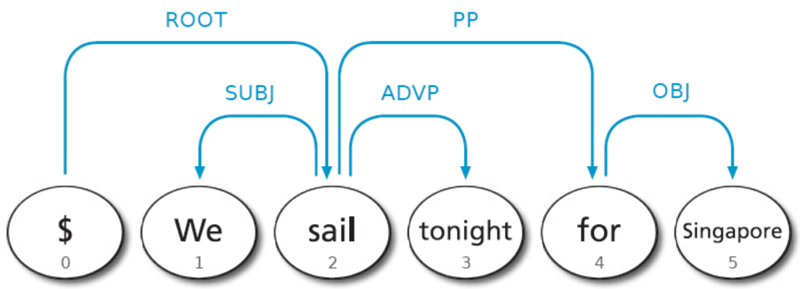
  - can be reduced to several binary decisions, but...
  - often better to handle multiple classes directly
  - how many classes? 2? 5? exponentially many?
- **Structured** prediction: many classes, strongly interdependent
  - Example: image segmentation (number of classes exponential to the # of segments)



$$\mathrm{x} = \begin{pmatrix} \mathrm{x}_{11} & \mathrm{x}_{12} & \cdots \\ \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots \\ \vdots & \vdots & \ldots \end{pmatrix} \Rightarrow \mathrm{y} = \begin{pmatrix} y_{11} & y_{12} & \cdots \\ y_{21} & y_{22} & \cdots \\ \vdots & \vdots & \ldots \end{pmatrix}$$

SAILING LAB

Example I: Dependency Parsing of Sentences



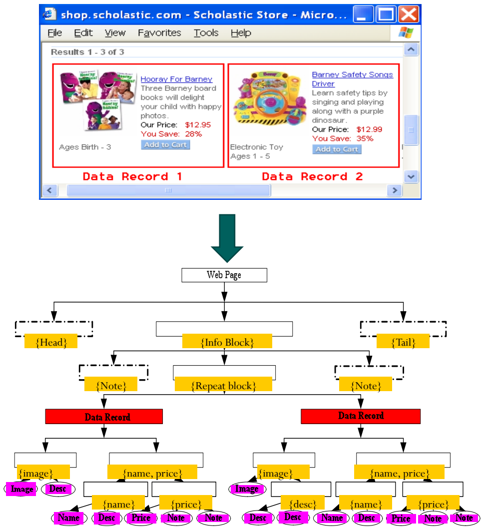
Challenge:
Structured outputs, and globally constrained to be a valid tree

SAILING LAB

Example II: Text Summarization

Australian novelist Peter Carey was awarded the coveted Booker Prize for fiction Tuesday night for his love story, "Oscar and Lucinda".
A panel of five judges unanimously announced the award of the $26,250 prize after an 80-minute deliberation during a banquet at London's ancient Guildhall.
The judges made their selection from 102 books published in Britain in the past 12 months and which they read in their homes.
Carey, who lives in Sydney with his wife and son, said in a brief speech that like the other five finalists he had been asked to attend with a short speech in his pocket in case he won.

Example III: Web-Data Extraction



Example IV: Topic Discovery/Extraction

⇒

"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.

Structured Prediction Graphical Models

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

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

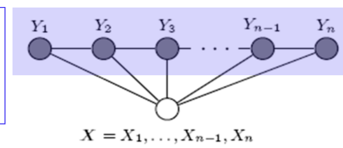
$$\mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \log \sum_{\mathbf{y}'} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}')) - \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})$$

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

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

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

$$F(\mathbf{x}, y; \mathbf{w}) = g(\mathbf{w}^\top \mathbf{f}(\mathbf{x}, y))$$



Structured Prediction Graphical Models

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

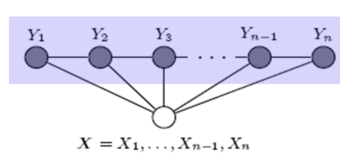
$$\mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \log \sum_{\mathbf{y}'} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}')) - \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})$$

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

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

Challenges:

- **SPARSE "Interpretable"** prediction model
- **Prior** information of structures
- **Latent** structures/variables
- **Time** series and non-stationarity
- **Scalable** to large-scale problems (e.g., 10^4 or larger input/output dimension)



SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegreative Genomics

Main Claims

- The sparse structures of natural language data (input) and of the NLP tasks (output) can be utilized to improve the **quality** of the solution and **interpretability** of the solution
- Over-parameterized models such as conventional NB/LR/SVM-style classifiers or parsers, topic models, or the related spectrum methods are not benefiting from sparse structures
- It is desirable to explore model spaces with structured sparsity for both **predictive models** (e.g., classifiers, parsers) and **explorative models** (e.g., topic models)

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegreative Genomics

Outline

- Sparse Structured Input-Output Models
  - … supervised learning
  - … convex optimization and log loss
  - … Frequentist-style shrinkage via regularization
- Sparse Topic Models
  - … unsupervised learning
  - … non-convex and likelihood-driven
  - … Bayesian-style posterior inference
- Sparse and Discriminative Topic Models?
  - … toward jointly explorative and predictive learning

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Basic text classification

Class Label		Word counts		Feature strength
1	=	0 learning 3 journal 2 intelligence 0 text 0 agent 1 internet 0 webwatcher 0 perl5 . . . 1 volume	x	?

y = X x β

EMNLP 2012

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Basic text classification

Class Label		Word counts		Feature strength
1	=	0 learning 3 journal 2 intelligence 0 text 0 agent 1 internet 0 webwatcher 0 perl5 . . . 1 volume	x	

$$\beta^* = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

Many non-zero coefficients:
Which words are truly significant?

EMNLP 2012

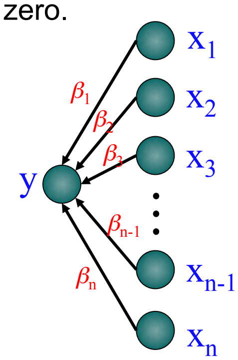
Sparsity: In a mathematical sense

- Consider least squares linear regression problem:
- Sparsity means most of the beta's are zero.

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\beta} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

subject to:

$$\sum_{j=1}^{p} \mathbb{I}[|\beta_j| > 0] \leq C$$



- But this is not convex!!! Many local optima, computationally intractable.

L1 Regularization (LASSO) (Tibshirani, 1996)

- A convex relaxation.

Constrained Form

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\beta} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

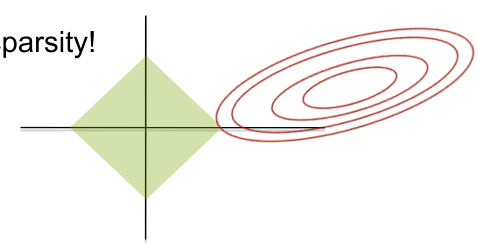
subject to:

$$\sum_{j=1}^{p} |\beta_j| \leq C$$

Lagrangian Form

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\beta} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1$$

- Still enforces sparsity!



SAILING LAB

Lasso for Sparse Regression

Class Label | Word counts | Feature strength

1 =

0	learning
3	journal
2	intelligence
0	text
0	agent
1	internet
0	webwatcher
0	perl5
.	
.	
.	
1	volume

x

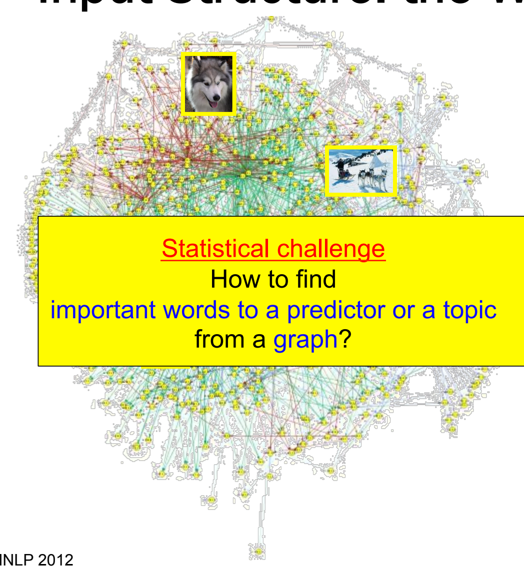
Lasso Penalty for sparsity

$$\beta^* = \arg\min_{\beta}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) + \lambda\sum_{j=1}^{J}|\beta_j|$$

Many zero associations (sparse results), but what if the problem has "structures"?

SAILING LAB

Input Structure: the WordNet



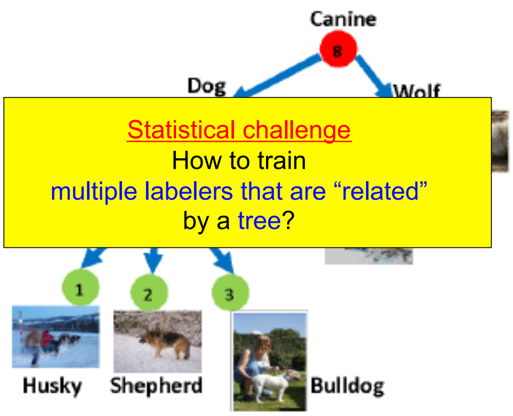
Statistical challenge
How to find
important words to a predictor or a topic
from a graph?

- The "network" of synsets in NL
  - Nodes (synsets) represent distinct concept
  - Links represent conceptual-semantic and lexical relations
  - Hidden knowledge and structure among concepts
  - Prior knowledge
  - Context

SAILING LAB

Output Structure: Task Hierarchy

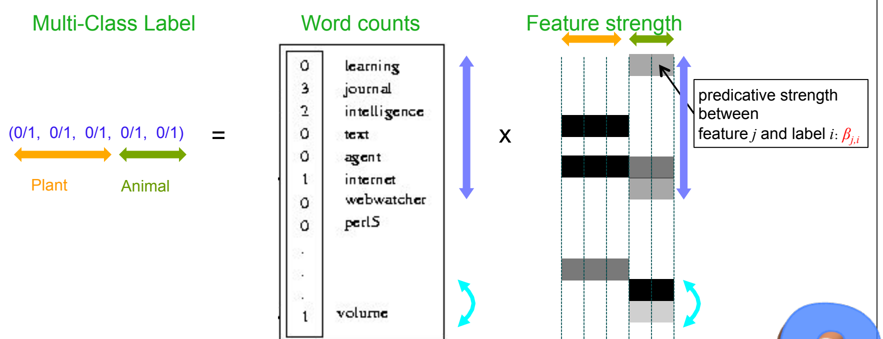
- E.g., the tree hierarchy in the DMOZ repository of the PASCAL Large Scale Hierarchical Text Classification challenge



Statistical challenge
How to train
multiple labelers that are "related"
by a tree?

SAILING LAB

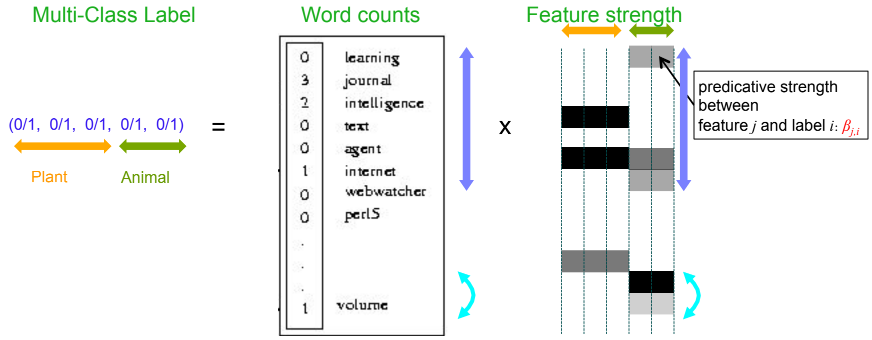
Sparse Structured Input/Output Lasso for Multi-task Learning



$$\beta^* = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{J} |\beta_j$$

How to combine information across multiple features/classes to increase the power?

Sparse Structured Input/Output Lasso for Multi-task Learning

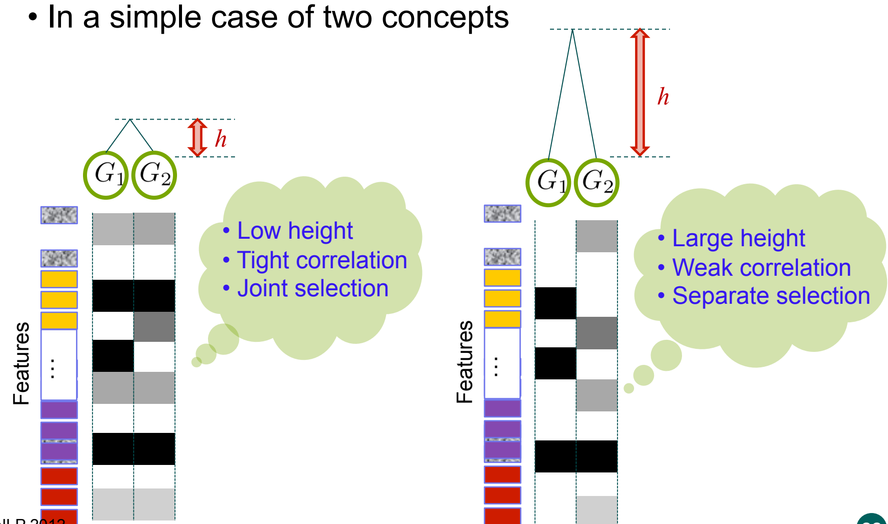


$$\beta^* = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^{J} |\beta_j|$$

+ We introduce
Structured fusion and/or group norm penalties

Tree-Guided Group Lasso

- In a simple case of two concepts

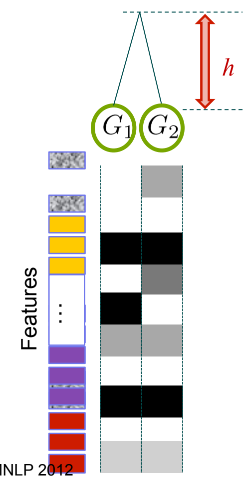


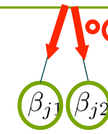
[Kim and Xing, ICML 2009]

SAILING LAB

Tree-Guided Group Lasso

- In a simple case of two concepts





Select the child nodes jointly or separately?

Tree-guided group lasso

$$\text{argmin } (y - X\beta)' \cdot (y - X\beta) + \lambda \sum_j \left[h\left(|\beta_{j1}| + |\beta_{j2}|\right) + (1-h)\left(\sqrt{\beta_{j1}^2 + \beta_{j2}^2}\right) \right]$$

L_1 penalty
- Lasso penalty
- Separate selection

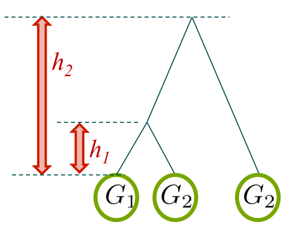
L_2 penalty
- Group lasso
- Joint selection

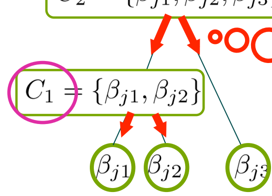
Elastic net

SAILING LAB

Tree-Guided Group Lasso

- For a general tree





Select the child nodes jointly or separately?

Tree-guided group lasso

$$\text{argmin } (y - X\beta)' \cdot (y - X\beta) + \lambda \sum_j \left[(1-h_2)\left(\sqrt{\beta_{j1}^2 + \beta_{j2}^2 + \beta_{j3}^2}\right) + h_2\left(|C_1| + |\beta_{j3}|\right) \right]$$

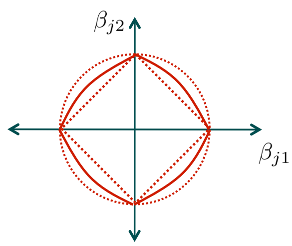
$$(1-h_1)\left(\sqrt{\beta_{j1}^2 + \beta_{j2}^2}\right) + h_1\left(|\beta_{j1}| + |\beta_{j2}|\right)$$

Joint selection — Separate selection

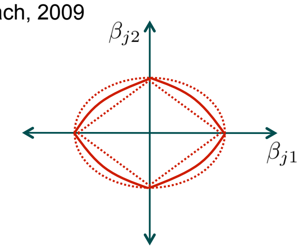
SAILING LAB

Proposition 1 *For each of the k-th output (gene), the sum of the weights w_v for all nodes $v \in V$ in T whose group G_v contains the k-th output (gene) as a member equals one. In other words, the following holds:*

$$\sum_{v:k \in G_v} w_v = \prod_{m \in Ancestors(v_k)} h_m + \sum_{l \in Ancestors(v_k)} (1 - h_l) \prod_{m \in Ancestors(v_l)} h_m = 1.$$

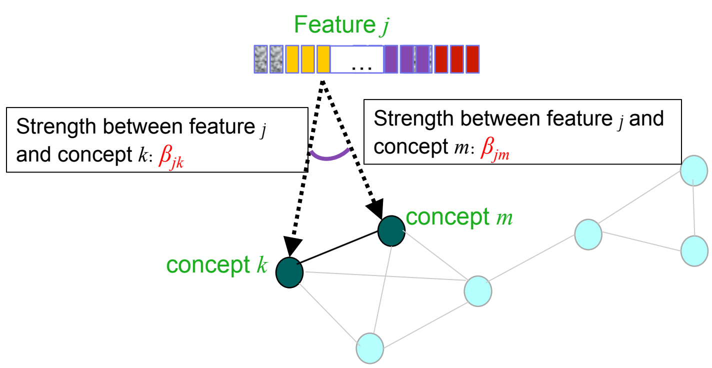


Previously, in Jenatton, Audibert & Bach, 2009



SAILING LAB

Graph-Guided Fused Lasso



- Fusion Penalty: $|\beta_{jk} - \beta_{jm}|$
- For two correlated concepts (connected in the network), the association strengths may have similar values.
  - Fusion effect propagates to the entire network
  - Association between features and subnetworks of concepts

[Kim and Xing, PLOS G 2009]

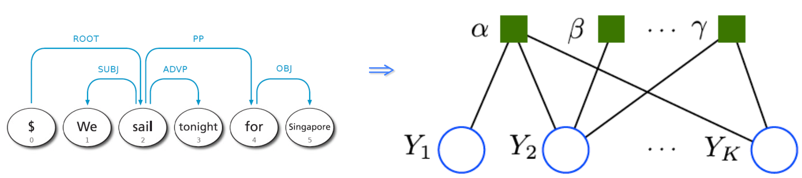
SAILING LAB

Full GM-based Loss Functions

$$y^\star = h(\mathrm{x}) \triangleq \arg\max_{y \in \mathcal{Y}} F(\mathrm{x}, y; \mathbf{w})$$

$$F(\mathrm{x}, y; \mathbf{w}) = g(\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y))$$

Represent **factorization** assumptions: $P(\mathbf{y}|x) = \frac{1}{Z} \prod_i \Psi_i(y_i) \prod_\alpha \Psi_\alpha(\mathbf{y}_\alpha)$



Inference: compute the MAP, marginals $\mu_i(y_i)$ and $\mu_\alpha(\mathbf{y}_\alpha)$, Z

- tractable when $\mathscr{G}$ is a tree, often intractable otherwise

SAILING LAB

Optimization

Original Problem: $\arg\max_\beta \equiv \mathcal{L}(\{\mathbf{x}_i, \mathbf{y}_i\}; \beta) + \Omega(\beta)$

Existing Methods:

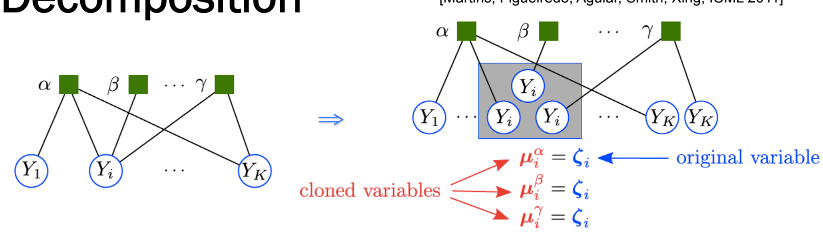
Interior-point Method (IPM) for Second-order Cone Programming (SOCP) or Quadratic Programming (QP)	2nd-order, computationally heavy	$\lambda \sum_{g \in \mathcal{G}} w_g \|\boldsymbol{\beta}_g\|_2 \Rightarrow \lambda \sum_{g \in \mathcal{G}} w_g t_g$ s.t. $\|\boldsymbol{\beta}_g\| \leq t_g$
Block Coordinate Descent	Cannot be easily be applied. Hard to compute the subgradient	Optimize $\boldsymbol{\beta}_g$ at one time

New Optimization Framework

- Main Difficulties:
  - Complex loss $\mathcal{L}(\{\mathbf{x}_i, \mathbf{y}_i\}; \beta)$, (e.g., GMs with intractable factors or loopy graphs)
    - Intractable inference
  - Complex shrinkage $\Omega(\beta)$, (e.g., overlapping group penalties)
    - Non-differentiable, non-separatable
- Our approaches:
  - Alternating Direction Dual Decomposition (AD^3) [Martins et al, ICML 2011]
  - Proximal Gradient [Chen et al, AOAS 2012]
  - Hierarchical Group Threshholding [Lee and Xing, 2012, submitted]
- Large number of training examples
  - Parallel computation
  - Map-Reduce on computing gradient
    - Map: calculate gradient on single example
    - Reduce: gather gradients computed by all map procedures, and calculate the sum
  - New multi-core framework ..

Alternating Directions Dual Decomposition

[Martins, Figueiredo, Aguiar, Smith, Xing, ICML 2011]



$$\textbf{maximize} \quad \overbrace{\boldsymbol{\theta}_\alpha \cdot \boldsymbol{\mu}_\alpha + \sum_{i \in \mathcal{N}(\alpha)} \left(\frac{\boldsymbol{\theta}_i}{|\mathcal{N}(i)|} + \boldsymbol{\lambda}_i^\alpha \right) \cdot \boldsymbol{\mu}_i^\alpha}^{\text{linear term}} - \overbrace{\frac{\eta}{2} \sum_{i \in \mathcal{N}(\alpha)} \|\boldsymbol{\mu}_i^\alpha - \boldsymbol{\zeta}_i^\alpha\|^2}^{\text{penalty term}}$$

$$\textbf{w.r.t.} \quad \boldsymbol{\mu}|_\alpha \in \textbf{MARG}(\mathcal{G}|_\alpha)$$

- Convergent to the primal and dual solutions (Glowinski and Le Tallec, 1989)
- O(1/ε) iterations suce for ε-accurate objective (He and Yuan, 2011)
- Solution is always sparse (only O(|N(α)|) nonzeros)
- Active set methods: seek the support of the solution by adding/removing components; very suitable for warm-starting (Nocedal and Wright, 1999)

Smooth Proximal Gradient Descent

[Chen et al and Xing, UAI 2011, AOAS 2012]

Original Problem:

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^J} f(\boldsymbol{\beta}) \equiv \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \Omega(\boldsymbol{\beta})$$

$$\Omega(\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha}\in\mathcal{Q}} \boldsymbol{\alpha}^T C\boldsymbol{\beta}$$

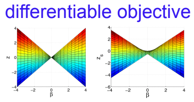
Separating overlapping constraints

Approximation Problem:

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^J} \widetilde{f}(\boldsymbol{\beta}) \equiv \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + f_\mu(\boldsymbol{\beta})$$

$$f_\mu(\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha}\in\mathcal{Q}} \boldsymbol{\alpha}^T C\boldsymbol{\beta} - \mu d(\boldsymbol{\alpha})$$

Smoothing non-differentiable objective



Gradient of the Approximation:

$$\nabla\widetilde{f}(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}-\mathbf{y}) + C^T\boldsymbol{\alpha}^*$$

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}\in\mathcal{Q}} \boldsymbol{\alpha}^T C\boldsymbol{\beta} - \mu d(\boldsymbol{\alpha})$$

$\nabla\widetilde{f}(\boldsymbol{\beta})$ is Lipschitz continuous with the Lipschitz constant L

$$L = \lambda_{\max}(\mathbf{X}^T\mathbf{X}) + L_\mu$$

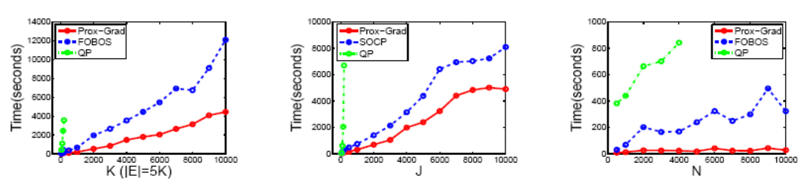
Convergence Rate

Theorem: If we require $f(\boldsymbol{\beta}^t) - f(\boldsymbol{\beta}^*) \le \epsilon$ and set $\mu = \frac{\epsilon}{2D}$, the number of iterations is upper bounded by:

$$t \le \sqrt{\frac{4\|\boldsymbol{\beta}^*\|_2^2}{\epsilon}\left(\lambda_{\max}(\mathbf{X}^T\mathbf{X}) + \frac{2D\|\Gamma\|^2}{\epsilon}\right)} = O(\frac{1}{\epsilon})$$

Remarks: state of the art IPM method for for SOCP converges at a rate $O(\frac{1}{\epsilon^2})$

Time complexity (Per-iteration): $O(J^2K + J\sum_{g\in\mathcal{G}}|g|)$ vs. $O\left(J^2(K+|\mathcal{G}|)^2(KN + J(\sum_{g\in\mathcal{G}}|g|))\right)$



(a) (b) (c)

What if the structure becomes too complex ?

- Too many groups in real problems:

$$\beta_{io-lasso} = \arg\min_{\beta} \sum_{k=1}^{K}\sum_{i=1}^{N}\left(Y_i^k - \sum_{j=1}^{p}\beta_j^k X_{ij} - \sum_{(r,s)\in U}\beta_{rs}^k Z_{i,rs}\right) + \lambda_1 \sum_{j=1}\sum_{k=1}\left|\beta_j^k\right|$$

$$+ \lambda_2 \sum_k \sum_m \sqrt{\sum_{(r,s)\in S_m} {\beta_{rs}^k}^2}$$

$$+ \lambda_3 \sum_j \sqrt{\sum_k {\beta_j^k}^2}$$

$$+ \lambda_4 \sum_k \sum_{(r,s)\in U}\left|\beta_{rs}^k\right|$$

Lasso penalty within group sparsity

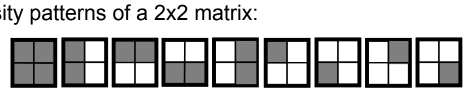
Easily hundreds of thousands of group constraints!

  - Recall that even SPG has a complexity of $O(J^2K + J\sum_{g\in\mathcal{G}}|g|)$
- And an optimization procedure must
  - minimize our objective function, and
  - induce correct sparsity patterns
- Hierarchical group-thresholding:
  - an algorithmic approach to directly reduce search space of sparsity, while optimizes the exact loss

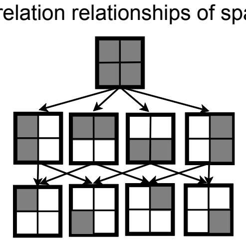
Hierarchical Group-Thresholding

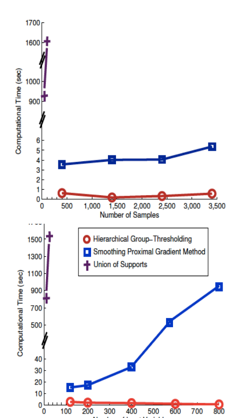
[Lee and Xing, Submitted 2012]

- DAG for Sparsity Patterns
  - All sparsity patterns of a 2x2 matrix:



  - A DAG of inclusion relation relationships of sparsity





- Hierarchical Group-Thresholding
  - Initialize B using ridge regression
  - Step 1: Traversing DAG, check the optimality condition of the zero pattern at each node. If the condition holds, set zero
  - Step 2: Update non-zero regression coefficients using coordinate descent

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegreative Genomics

What if the function is non-linear?

Group Sparse Additive Models [Ying, Chen, Xing, ICML 2012]

- Assume G is a partition of {1, · · · , p}, i.e., the groups in G do not overlap.
- The optimization problem is
- $$\min_{\mathbf{f}} L(\mathbf{f}) + \lambda\Omega_{\text{group}}(\mathbf{f}),$$

  where

  $$\Omega_{\text{group}}(\mathbf{f}) = \sum_{g\in\mathcal{G}} \sqrt{|g|}\|\mathbf{f}_g\| = \sum_{g\in\mathcal{G}} \sqrt{|g|}\sqrt{\sum_{j\in g}\mathbb{E}\left[f_j^2(X_j)\right]}.$$
- Non-trivial to solve due to
  - correlation structure of component functions within the group
  - non-smoothness of functional group penalty

EMNLP 2012

35

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegreative Genomics

Toward Human-Level Intelligence

- Now we have dealt with high feature dimension
  - Sparsity
- and we have know how to leverage structural knowledge
  - Structured shrinkage
- What about massive **concept space**?

The IMAGENET knowledge ontology

Data: >20 million images

Features: ~1 million (number comes from the top performing system in ILSVRC10, [Lin et al. 2011])

Classes: ~22k classes

EMNLP 2012

36

Output Coding

M	1	2	3	...	...	K
1	1	1	1	0	0	0
2	-1	0	0	1	1	0
...	0	-1	0	-1	0	1
C	0	0	-1	0	-1	-1

- Every class is now represented by a bit-string
  - ***Coding***: a codeword is assigned to each class
  - ***Decoding***: given test data, look for most similar class codeword
- Predict bit by bit through binary or ternary classifier – this is much easier than the 1 vs C-1 classifier
- Decoding the bit-string – error correcting

[Zhao and Xing, 2012, submitted]

Learning the Coding Matrix

M	1	2	3	...	...	K
1	1	1	1	0	0	0
2	-1	0	0	1	1	0
...	0	-1	0	-1	0	1
C	0	0	-1	0	-1	-1

- Accuracy of base binary classifiers for bit-prediction
  - Use category hierarchy for a measure of separability
  - Large intra-partite similarity + small inter-partite similarity
- Strong error-correcting ability
  - Maximize distance between rows of coding matrix
- Fault tolerance
  - Introduction of ignored classes: {-1,0,+1} instead of {-1,+1}

$$\max_{\mathbf{B}} \quad F_b(\mathbf{B}) - \lambda_r F_r(\mathbf{B}) - \lambda_c \sum_{l=1}^{L} ||\beta_l||_2^2$$

$$s.t. \quad \mathbf{B} \in \{-1, 0, +1\}^{K \times L}$$

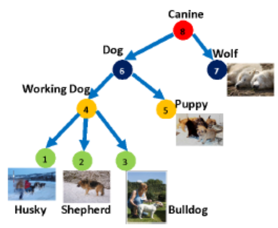
$$\sum_{k=1}^{K} (|B_{kl}| + B_{kl}) \geq 2, \ \forall l = 1, \ldots, L$$

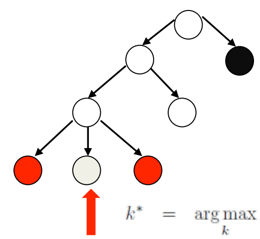
$$\sum_{k=1}^{K} (|B_{kl}| - B_{kl}) \geq 2, \ \forall l = 1, \ldots, L$$

$$\sum_{l=1}^{L} |B_{kl}| \geq 1, \ \forall k = 1, \ldots, K$$

Probabilistic Decoding

- Output code can have real semantic meaning
  - E.g., encoding a tree path in a label taxonomy
- Probabilistic decoding:
  - bit i depend on bit j probabilistically
- Define prior $P(y_l|\hat{y} = k)$ using tree hierarchy
  - Graph coloring: all nodes participating in i-th bit prediction are colored (red for positive, black for negative)
  - Task: what is the probability of node *k* being colored red?





$$k^* = \arg\max_k P(\hat{y} = k | w_1, \ldots, w_L, \ldots)$$

EMNLP 2012

39

Multi-Way Classification Accuracy

- DMOZ repository in PASCAL Hierarchical Text Classification challenge

Data set	# classes	# training data	# test data	# features
DMOZ-small	1139	6323	1858	1199848
DMOZ-large	12294	93805	34905	1199856

	DMOZ small		DMOZ large	
Algorithm	Top 1	Top 5	Top 1	Top 5
OVR	50.91	64.72	37.24	44.87
RDOC	41.77	54.52	5.13	7.99
RSOC	42.30	58.40	5.47	7.23
SpectralOC	44.83	59.10	22.10	23.82
SSOC	**56.67**	**67.33**	**41.28**	**46.71**

EMNLP 2012

[Zhao and Xing, Submitted 2012]

40

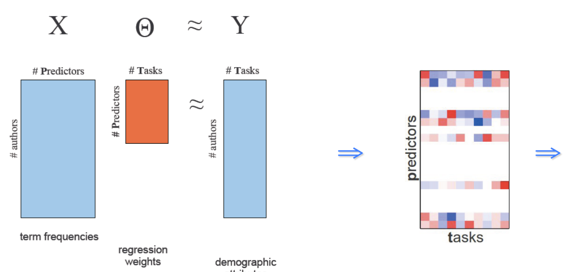
Discovering Sociolinguistic Associations on Twitter

- Twitter Gardenhose feed from March 1-7, 2010
- 9250 authors, 380,000 messages, 4.7 million tokens
- Filters:
  - At least 20 messages (in Gardenhose)
  - Messages must include GPS within a USA zipcode
  - No more than 1000 followers, followees
- GPS → Zipcode → U.S. Census Demographic Statistics
  - Zipcodes commonly proxy for demographics in public health.
  - Careful! Twitter users are not an unbiased sample from a zipcode.

+ United States Census 2010

[Eisenstein, Smith, and Xing. ACL 2011]

Demographic multi-prediction



$L1/L\infty$ regularizer

$$\Omega(\Theta) = \sum_t \max_p |\theta_{pt}|$$

aka **multi-output lasso** (Turlach et al 2005)

vocabulary	# features	average	white	Afr. Am.	Hisp.		Eng. lang.	Span. lang.	other lang.		urban	family	renter	med. inc.
full	5418	0.260	0.337	0.318	0.296		0.384	0.296	0.256		0.155	0.113	0.295	0.152
multi-output lasso	394.6	0.260	0.326	0.308	0.304		0.383	0.303	0.249		0.153	0.113	0.302	0.156
SVD		0.237	0.321	0.299	0.269	···	0.352	0.272	0.226	···	0.138	0.081	0.278	0.136
highest variance		0.220	0.309	0.287	0.245		0.315	0.248	0.199		0.132	0.085	0.250	0.135
most frequent		0.204	0.294	0.264	0.222		0.293	0.229	0.178		0.129	0.073	0.228	0.126

[Eisenstein, Smith, and Xing. ACL 2011]

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Sociolinguistic Associations

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
as			-	+	-					
awesome	+	-					-		-	+
break			-	+	-	-				
campus			-	+	-	-				
dead	-	+		-	+		+		+	
hell			-	+	-	-				
shit	-								+	
train				-	+				+	
will			-	+	-					
would				+					-	

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
-_-	-		+	-	+	+	+			
;)		-	+	-	+					
:(		-								
:)		-								
:d	+	-	+	-	+					

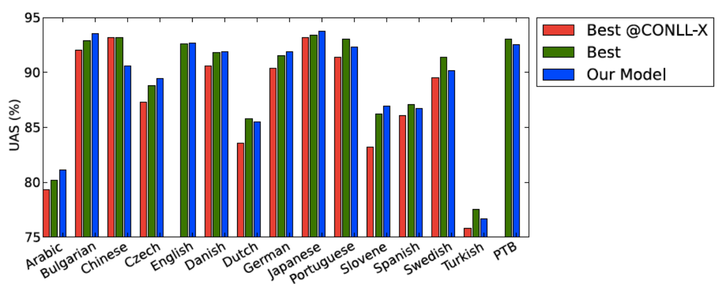
	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
bbm	-	+		-		+	+		+	
lls		+	-	+	-	-				
lmaoo	-	+	+	-	+	+	+		+	
lmaooo	-	+	+	-	+	+	+		+	
lmaoooo	-	+	+	-	+	+			+	
lmfaoo	-		+	-	+	+			+	
lmfaooo	-		+	-	+	+			+	
lml	-	+	+	-	+	+	+		+	-
odee	-		+	-	+		+		+	
omw	-	+	+	-	+	+	+		+	
smfh	-	+	+	-	+	+	+		+	
smh	-	+					+		+	
w|	-		+	-	+	+	+		+	

[Eisenstein, Smith, and Xing. ACL 2011]

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Dependency Parsing

Datasets from CoNLL-2006 and CoNLL-2008 shared tasks



- Best published includes:
  - **transition-based** (Nivre et al., 2006; Huang and Sagae, 2010),
  - **graph-based** (McDonald et al., 2006; Koo and Collins, 2010),
  - **hybrid** (Nivre and McDonald, 2008; Martins et al., 2008a),
  - **turbo parsers** (Martins et al., 2010; Koo et al., 2010)

[Martins, Smith, and Xing. ACL 2009, Martins, EMNLP 2010, 2011]

Outline

- Sparse Structured Input-Output Models
  - … supervised learning
  - … convex optimization and log loss
  - … Frequentist-style shrinkage via regularization
- Sparse Topic Models
  - … unsupervised learning
  - … non-convex and likelihood-driven
  - … Bayesian-style posterior inference
- Sparse and Discriminative Topic Models?
  - … toward jointly explorative and predictive learning

Modeling Semantics: e.g., Topic Models

Generating a document

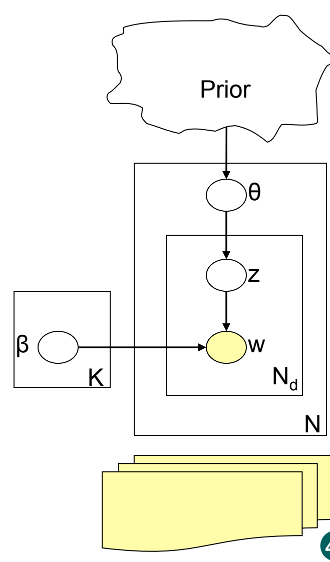
– *Draw* θ from the prior

For each word n

- Draw z_n from $multinomial(\theta)$

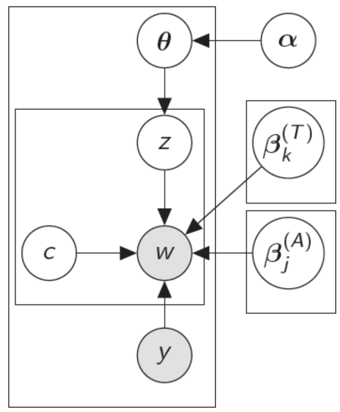
- Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from $multinomial(\beta_{z_n})$

- **Prior over topic Vector**
  - Latent Dirichlet Allocation (LDA)
  - Correlated priors (CTM)
  - Hierarchical priors
- **Topics**
  - Unigram, bigrams, etc
- **Document structure**
  - Bag of words
  - Multi-modal
  - Side information

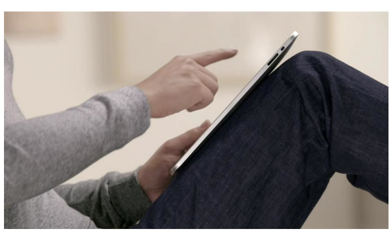


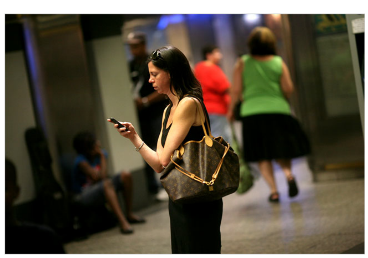
Modeling and Inference Complexity

- What if we want to combine latent topics with additional facets, such as geography in a unsupervised fashion?
- Additional latent variables decide which facet is responsible for each token (e.g. Ahmed and Xing 2010).
- That's twice as many latent variables per document!

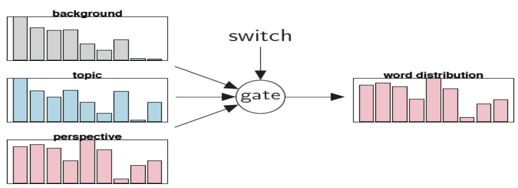


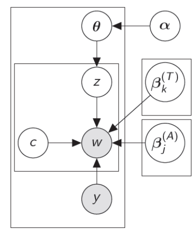
Compact modes needed on mobile devices



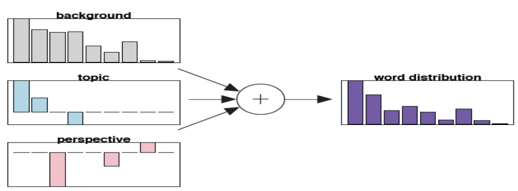


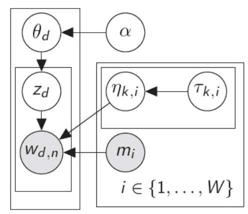
Sparse Additive Generative Models





versus

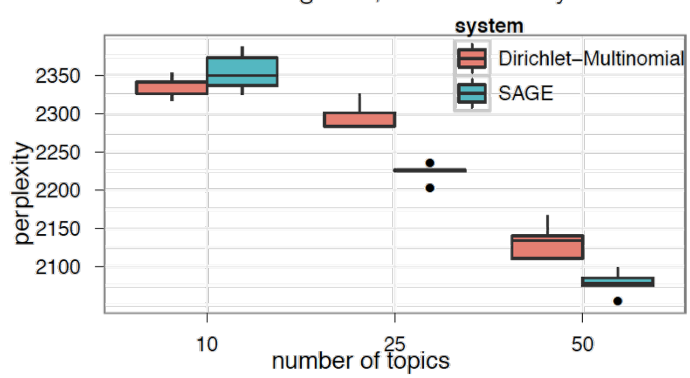




J. Eisenstein, A. Ahmed, E.P. Xing. ICML, 2011

Model Compression on Text

- NIPS dataset: 1986 training docs, 10K vocabulary

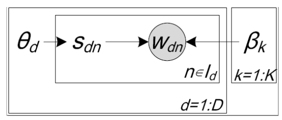


- Adaptive sparsity:
  - 5% non-zeros for 10 topics
  - 1% non-zeros for 50 topics

SAILING LAB

Sparse Topical Coding

- Goal: design a non-probabilistic topic model that is amenable to
  - direct control on the posterior sparsity of inferred representations
  - avoid dealing with normalization constant when considering supervision or rich features
  - seamless integration with a convex loss function (e.g., svm hinge loss)
- We extend sparse coding to hierarchical sparse topical coding
  - word code θ
  - document code $\boldsymbol{s}$



reconstruction loss — sparse codes

$$\min_{\{\boldsymbol{\theta}_d, \mathbf{s}_d\},\boldsymbol{\beta}} \sum_{d,n\in I_d} \ell(w_{dn}, \mathbf{s}_{dn}^\top \boldsymbol{\beta}_{.n}) + \lambda \sum_d \|\boldsymbol{\theta}_d\|_1 + \sum_{d,n\in I_d} (\gamma \|\mathbf{s}_{dn} - \boldsymbol{\theta}_d\|_2^2 + \rho \|\mathbf{s}_{dn}\|_1)$$

$$\text{s.t.}: \boldsymbol{\theta}_d \geq 0,\ \mathbf{s}_{dn} \geq 0,\ \forall d, n \in I_d;\ \ \boldsymbol{\beta}_k \in \mathcal{P},\ \forall k,$$

truncated aggregation

non-negative codes — topical bases

EMNLP 2012 — J. Zhu, & E.P. Xing. UAI, 2011 — 51

SAILING LAB

Algorithms on Sparse Latent Space Models

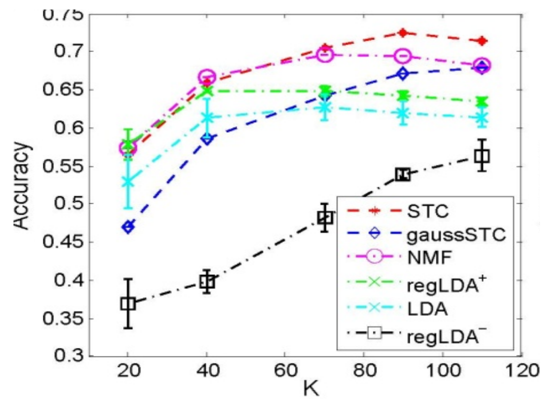
- Complex objective
  - Non-convex, but often bi-convex
  - Often additional non-negativity constraints other than sparsity
- Hierarchical sparse coding
  - Greedy algorithm for the non-convex L_0 “pseudo-norm”:
    - select the element with maximum correlation with the residual
    - known as “matching pursuit” (Mallat & Zhang, 1993)
  - For the convex L_1 norm, many algorithms:
    - Soft-thresholding with coordinate descent (Friedman et al., 2007; Zhu & Xing, 2011)
    - Proximal methods (Nesterov, 2007; Jenatton et al., 2010, Chen et al 2011)
    - Active-set methods (Roth & Fischer, 2008)
    - Online/stochastic variants
    - …
- Dictionary (topic) learning
  - projected gradient descent
  - any faster alternative method can be used

EMNLP 2012 — 52

Comparisons

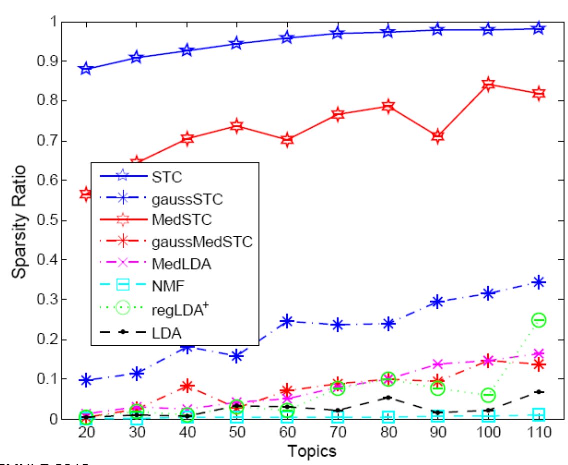
LDA vs. STC

[Zhu and Xing, UAI 2011]



Sparse word codes

- Sparsity ratio: percentage of zeros



- NMF: non-negative matrix factorization
- MedLDA (Zhu et al., 2009)
- regLDA: LDA with entropic regularizer
- gaussSTC: use L2 rather than L1-norm

Outline

- Sparse Structured Input-Output Models
  - … supervised learning
  - … convex optimization and log loss
  - … Frequentist-style shrinkage via regularization
- Sparse Topic Models
  - … unsupervised learning
  - … non-convex and likelihood-driven
  - … Bayesian-style posterior inference
- Sparse and Discriminative Topic Models?
  - … toward jointly explorative and predictive learning

Predictive Subspace Learning with *Supervision*

- Unsupervised latent subspace representations are generic but can be sub-optimal for predictions
- Many datasets are available with supervised side information

- Tripadvisor Hotel Review (http://www.tripadvisor.com)

"Lovely welcomming staff, good rooms that give a good nights sleep, downtown location"
Meramees Hostel
Save Review
SheikhSahib 10 contributions
London
Jul 7, 2009 | Trip type: Friends getaway
This hotel is just of the side streets of Talat Harb, one of the main arteries to downtown Cairo. It is walking distance to the Nile, riverfront hotels, Egyptian Museum, and there are many eateries in the area at night when it is still bustling. Only a short cab ride away from the Old Fatimid Cairo.

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

My ratings for this hotel
Value
Rooms
Location
Cleanliness
Service

Date of stay February 2009
Visit was for Leisure
Traveled with With Friends
Member since July 03, 2009
Would you recommend this hotel to a friend? Yes

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

- Many others

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

IMAGENET

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

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegreative Genomics

MLE versus Max-margin Learning

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

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

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

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

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

EMNLP 2012 57

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegreative Genomics

MaxEnt Discrimination Markov Network

(Zhu et al, ICML 2008, Zhu and Xing, JMLR 2009)

- Structured MaxEnt Discrimination (SMED):

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

generalized maximum entropy or *regularized* KL-divergence

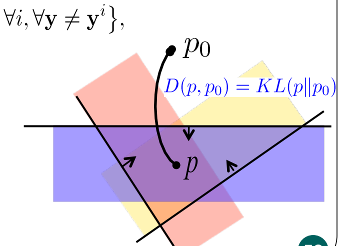
- Feasible subspace of weight distribution:

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

expected margin constraints.

- Average from a distribution of M³Ns

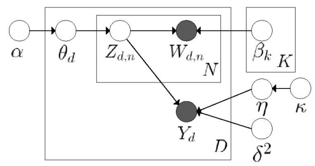
$$h_1(\mathbf{x}; p(\mathbf{w})) = \arg \max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \int p(\mathbf{w}) F(\mathbf{x}, \mathbf{y}; \mathbf{w}) d\mathbf{w}$$



EMNLP 2012 58

Maximum Entropy Discrimination LDA (MedLDA)

- Bayesian sLDA:



- MED Estimation:
  - MedLDA Regression Model

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

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

model fitting

predictive accuracy

  - MedLDA Classification Model

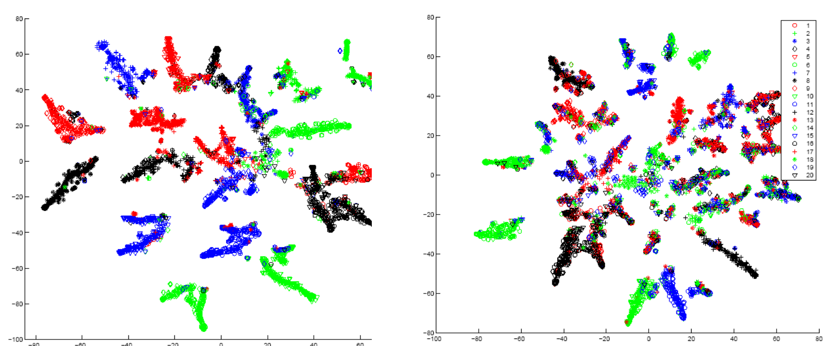
$$\mathrm{P2(MedLDA}^c): \min_{q,q(\eta),\alpha,\beta,\xi} \mathcal{L}(q) + C\sum_{d=1}^{D}\xi_d$$

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

(Zhu et al, ICML 2009, JMLR 2012)

Document Modeling

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

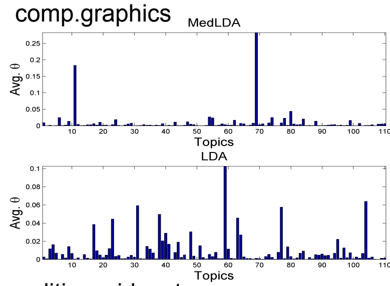


MedLDA LDA

SAILING LAB

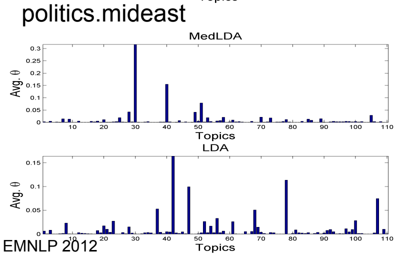
Document Modeling

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



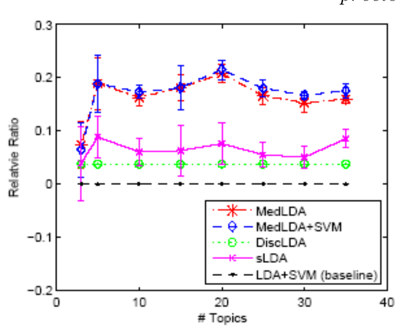
T 30	T 40	T 51	T 42	T 78	T 47
israel	turkish	israel	israel	jews	armenian
israeli	armenian	lebanese	israeli	jewish	turkish
jews	armenians	israeli	peace	israel	armenians
arab	armenia	lebanon	writes	israeli	armenia
writes	people	people	article	arab	turks
people	turks	attacks	arab	people	genocide
article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

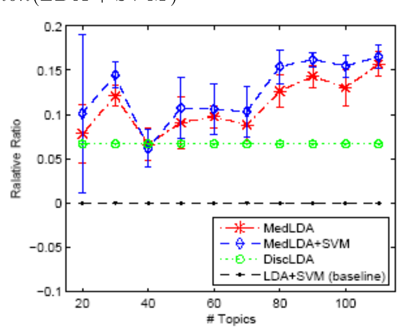
SAILING LAB

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

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



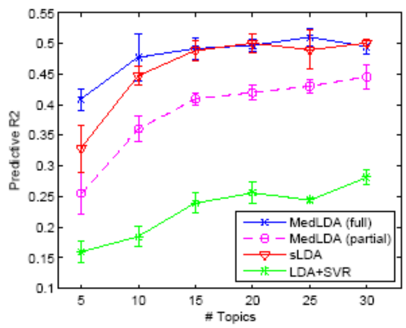


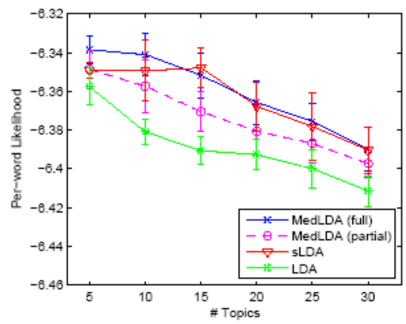
SAILING LAB

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

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

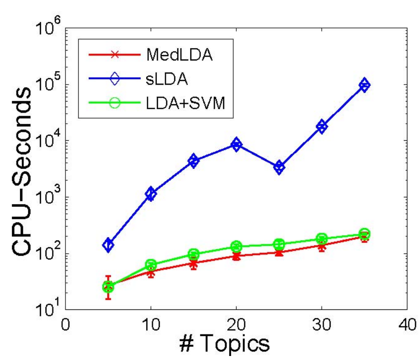




SAILING LAB

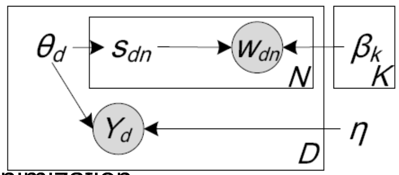
Time Efficiency

- Binary Classification



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

Supervised STC



- Joint loss minimization

$$\min_{\{\theta_d\},\{\mathbf{s}_d\},\beta,\eta} \quad f(\{\theta_d\},\{\mathbf{s}_d\},\beta) + C\mathcal{R}_h(\{\theta_d\},\eta) + \frac{1}{2}\|\eta\|_2^2$$

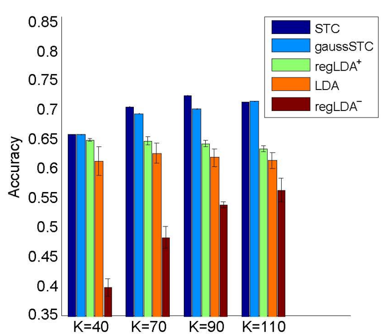
$$\text{s.t.}: \quad \theta_d \geq 0,\ \forall d;\ \mathbf{s}_{dn} \geq 0,\ \forall d, n \in I_d;\ \beta_k \in \mathcal{P},\ \forall k,$$

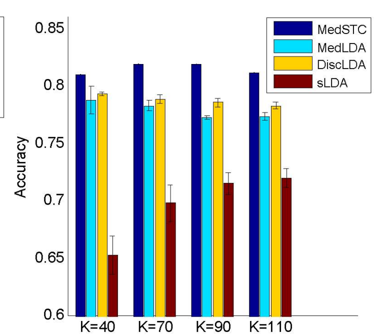
  - coordinate descent alg. applies with closed-form update rules
  - No sum-exp function; seamless integration with non-probabilistic large-margin principle

(Zhu and Xing, UAI 2011)

Classification accuracy

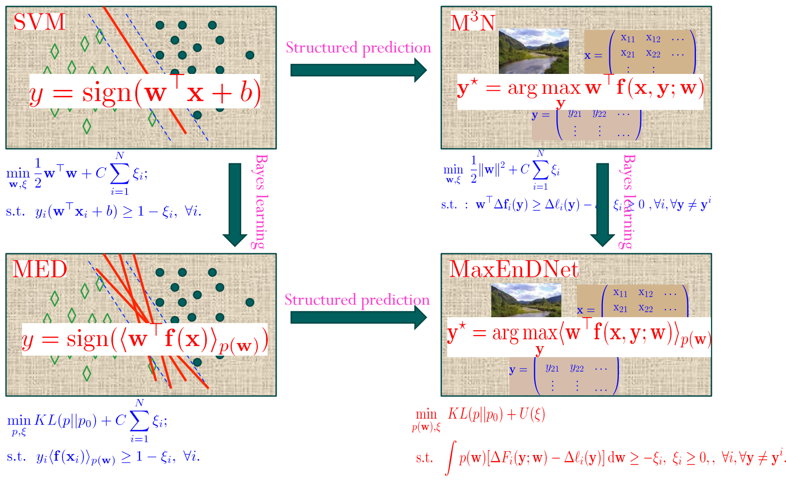
- 20 newsgroup data:





SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Summary: Margin-based Learning Paradigms



EMNLP 2012

67

SAILING LAB
Laboratory for Statistical Artificial InteLligence & INtegrative Genomics

Conclusion and Challenges: Learning Sparse Structured Input/Output

- **Model:** Multi-Task Classifiers, Graphical Models, Topic Models
- **Algorithm:** Convex and non-convex optimization
- **Representation:** Concept encoding/decoding
- **System:** multi-core, distributed file system, shared memory, cloud
- **Theory:** convergence, sample complexity, asymptotic consistency/sparsistency, error bounds, etc

EMNLP 2012

68

Thanks!

Reference: