Probabilistic Graphical Models

RegBayes: a general paradigm for learning GMs

Eric Xing
(courtesy to Jun Zhu)
Lecture 29, May 1, 2013

Reading:
© Eric Xing @ CMU, 2005-2013
Bayesian Inference

- A coherent framework of dealing with uncertainties

\[ p(M|x) = \frac{p(x|M)\pi(M)}{\int p(x|M)\pi(M)dM} \]

- \( M \): a model from some hypothesis space

- \( x \): observed data

Bayes’ rule offers a mathematically rigorous computational mechanism for combining prior knowledge with incoming evidence

Thomas Bayes (1702 – 1761)

---

Parametric Bayesian Inference

- \( \mathcal{M} \) is represented as a finite set of parameters \( \theta \)

- A parametric likelihood: \( x \sim p(.|\theta) \)

- Prior on \( \theta \): \( \pi(\theta) \)

- Posterior distribution

\[ p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)d\theta} \propto p(x|\theta)\pi(\theta) \]

Examples:

- Gaussian distribution prior + 2D Gaussian likelihood \( \rightarrow \) Gaussian posterior distribution
- Dirichlet distribution prior + 2D Multinomial likelihood \( \rightarrow \) Dirichlet posterior distribution
- Sparsity-inducing priors + some likelihood models \( \rightarrow \) Sparse Bayesian inference

© Eric Xing @ CMU, 2005-2013
Nonparametric Bayesian Inference

\( \mathcal{M} \) is a richer model, e.g., with an infinite set of parameters

- A nonparametric likelihood: \( x \sim p(\cdot|\mathcal{M}) \)
- Prior on: \( \mathcal{M} \sim \pi(\mathcal{M}) \)
- Posterior distribution:

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

Examples:
- see next slide
Why Bayesian Nonparametrics?

- Let the data speak for themselves
- Bypass the model selection problem
  - let data determine model complexity (e.g., the number of components in mixture models)
  - allow model complexity to grow as more data observed

Can we further control the posterior distributions?

It is desirable to further regularize the posterior distribution

- An extra freedom to perform Bayesian inference
- Arguably more direct to control the behavior of models
- Can be easier and more natural in some examples
Can we further control the posterior distributions?

- Directly control the posterior distributions?
  - Not obvious how ...

**Hard constraints**
(A single feasible space)

**Soft constraints**
(many feasible subspaces with different complexities/penalties)

---

A reformulation of Bayesian inference

- Bayes’ rule is equivalent to:

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

**A direct but trivial constraint on the posterior distribution**

E.T. Jaynes (1988): “this fresh interpretation of Bayes’ theorem could make the use of Bayesian methods more attractive and widespread, and stimulate new developments in the general theory of inference”
Regularized Bayesian Inference

\[
\inf_{q(M), \xi} \text{KL}(q(M)\|\pi(M)) - \int_{\mathcal{M}} \log p(D|M)q(M)dM + U(\xi)
\]

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

where, e.x.,

\[\mathcal{P}_{\text{post}}(\xi) \overset{\text{def}}{=} \left\{ q(M) | \forall t = 1, \ldots, T, \ h(Eq(\psi; D)) \leq \xi_t \right\} ,\]

and

\[U(\xi) - \sum_{t=1}^{T} I(\xi_t - \gamma) - I(\xi - \gamma)\]

Solving such constrained optimization problem needs convex duality theory

So, where does the constraints come from?

Recall our evolution of the Max-Margin Learning Paradigms

SVM
\[y = \text{sign}(w^T x + b)\]
\[\min_{w, b} \frac{1}{2}||w||^2 + C \sum_{i} \zeta_i \quad y'(w^T x + b) \geq 1 - \zeta_i \quad \forall i\]

MED
\[y = \text{sign}(f(x, w)q(w))\]
\[\min_{Q} \text{KL}(Q\|Q_0) \quad y'(f(x, w))q(w) \geq \zeta_i \quad \forall i\]

M3N
\[y = \arg \max_{y \in Y(x)} F(x, y, w)\]
\[\min_{w, \zeta_i} \frac{1}{2}||w||^2 + C \sum_{i} \zeta_i \quad w^T[f(x') - f(x, y)] \geq (y', y) - \zeta_i \quad \forall i. \forall y \neq y'\]

MED-MN
\[= \text{SMED + “Bayesian” M3N}\]
Maximum Entropy Discrimination
Markov Networks

- Structured MaxEnt Discrimination (SMED):

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

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

- Feasible subspace of weight distribution:

\[
\mathcal{F}_1 = \{p(w) : \int p(w)[\Delta F_i(y; w) - \Delta t_i(y)] dw \geq -\xi_i, \forall i, \forall y \neq y^i\}
\]

- Average from distribution of M³Ns

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

Can we use this scheme to learn models other than MN?
Recall the 3 advantages of MEDN

- An averaging Model: PAC-Bayesian prediction error guarantee (Theorem 3)
  \[ \Pr_Q(M(h, x, y) \leq 0) \leq \Pr_P(M(h, x, y) \leq \gamma) + O\left(\frac{\gamma^{-2}KL(p||q) \ln(N)}{N}\right) \]

- Entropy regularization: Introducing useful biases
  - Standard Normal prior => reduction to standard M3N (we've seen it)
  - Laplace prior => Posterior shrinkage effects (sparse M3N)
    \[ \min_{\alpha} \frac{1}{2} \sum \left( x_i^2 + 1 \right) \log x_i + \frac{1}{2} \log \left( \frac{1}{2} \right) + C \sum_{i=1}^{N} \left( 1 + \frac{1}{\sqrt{\alpha}} \right) \left( x_i^2 + 1 \right) \]
  - PoMEN

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

Latent Hierarchical MaxEnDNet

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

- Hierarchical labeling
  - Advantages:
    - Computational efficiency
    - Long-range dependency
    - Joint extraction
Partially Observed MaxEnDNet (PoMEN) (Zhu et al, NIPS 2008)

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

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

  \[ \begin{align*} 
  \text{P2(PoMEN)}: \quad & \min_{p(w(z))} \sum_{i} K(p(w(z)) || p_0(w(z))) + U(\xi) \\
  & \text{s.t. } p(w(z)) \in \mathcal{F}_2, \xi_i \geq 0, \forall i. 
  \end{align*} \]

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

- Prediction: \( h_2(x) = \arg \max_{y \in Y(x)} \sum_{i} \int p(w,z) F(x,y,z;w) \ dw \)

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 \( p(w) \)

    \[ \begin{align*} 
    & \min_{p(w) \in \mathcal{F}_1} \sum_{i} K(p(w)||p_0(w)) + C \xi_i \\
    & \text{s.t. } p(w) \in \mathcal{F}_1, \xi_i \geq 0, \forall i. 
    \end{align*} \]

  \[ \mathcal{F}_1 = \{ p(w) : \int p(w)[\Delta F(y,x;w) - \Delta F(y;y;w)] \ dw \geq -\xi, \forall i, \forall y \} \]

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

    \[ \begin{align*} 
    & \min_{p(z) \in \mathcal{F}_1} \sum_{i} K(p(z)||p_0(z)) + C \xi_i \\
    & \text{s.t. } p(z) \in \mathcal{F}_1, \xi_i > 0. 
    \end{align*} \]

  \[ \mathcal{F}_1 = \{ p(z) : \int p(z)[\Delta F(y,x;w) - \Delta F(y;y;w)] \ dw \geq -\xi, \forall i, \forall y \} \]

  Equivalently reduced to an LP with a polynomial number of constraints
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

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:
Page-Level Evaluations

- Supervision Level 1:
  - Leaf nodes and data record nodes are labeled

- Supervision Level 2:
  - Level 1 + the nodes above data record nodes

Key message from PoMEN

- Structured MaxEnt Discrimination (SMED):
  \[
  P_1 : \min_{p(w,z)} \int KL(p(w,z)||p_0(w,z)) + \sum_i \xi_i, \quad \text{s.t. } p(w,z) \in \mathcal{F}_1, \xi_i \geq 0, \forall i.
  \]
  
  generalized maximum entropy or regularized KL-divergence

- Feasible subspace of weight distribution:
  \[
  \mathcal{F} = \{p(w,z) : \int \int p(w,z) [\Delta E_i(y;w,z) - \Delta E_i(y')] \, dw\,dz \geq -\xi_i, \forall i, \forall y \neq y'\},
  \]
  
  expected margin constraints

- Average from distribution of PoMENs
  \[
  h(x) = \arg\max_{y \in \Theta(x)} \int \int p(w,z) F(x,y,z;w) \, dw\,dz
  \]

- We can use this for any \( p \) and \( p_0 \)!
An all inclusive paradigm for learning general GM --- RegBayes

\[
\inf_{q(M), \xi} \text{KL}(q(M)||\pi(M)) - \int_M \log p(D|M)q(M)dM + U(\xi)
\]
\[\text{s.t. : } q(M) \in \mathcal{P}_{\text{post}}(\xi),\]

Predictive Latent Subspace Learning
via a large-margin approach

... where M is any subspace model and p is a parametric Bayesian prior
Unsupervised Latent Subspace Discovery

- Finding latent subspace representations (an old topic)
  - Mapping a high-dimensional representation into a latent low-dimensional representation, where each dimension can have some interpretable meaning, e.g., a semantic topic

- Examples:
  - Topic models (aka LDA) [Blei et al 2003]
  - Total scene latent space models [Li et al 2009]
  - Multi-view latent Markov models [Xing et al 2005]
  - PCA, CCA, ...

Predictive Subspace Learning with Supervision

- Unsupervised latent subspace representations are generic but can be sub-optimal for predictions
- Many datasets are available with supervised side information
  - Tripadvisor Hotel Review (http://www.tripadvisor.com)
  - LabelMe http://labelme.csail.mit.edu/
  - Flickr (http://www.flickr.com/)
  - Many others

- Can be noisy, but not random noise (Ames & Naaman, 2007)
  - labels & rating scores are usually assigned based on some intrinsic property of the data
  - helpful to suppress noise and capture the most useful aspects of the data

- Goals:
  - Discover latent subspace representations that are both predictive and interpretable by exploring weak supervision information
I. LDA: Latent Dirichlet Allocation

(Blei et al., 2003)

\[ p(\theta, z, W | \alpha, \beta) = \prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \]

- Joint Distribution: \[ p(\theta, z, W | \alpha, \beta) \]
- Variational Inference with \( q(z, \theta) \sim p(z, \theta | W, \alpha, \beta) \)

\[ \mathcal{L}(q) \equiv -E_q[\log p(\theta, z, W | \alpha, \beta)] - H(q(z, \theta)) > - \log p(W | \alpha, \beta) \]
- Minimize the variational bound to estimate parameters and infer the posterior distribution

---

Maximum Entropy Discrimination LDA (MedLDA)

(Zhu et al, ICML 2009)

- Bayesian sLDA:

- MED Estimation:
  - MedLDA Regression Model
    \[ P_1(\text{MedLDA}^\ast) : \min_{q, \alpha, \beta, \xi, \zeta} \mathcal{L}(q) + C \sum_{d=1}^{D} (\xi_d + \zeta_d) \]
    s.t. \( \forall d \): \[ 0 \leq y_d + E[q] \rightarrow Z_d \leq \xi_d + \zeta_d, \]
    \[ 0 \leq -y_d + E[q] \rightarrow Z_d \leq \xi_d + \zeta_d, \]
    \[ \xi_d \geq 0, \zeta_d \geq 0, v_d \]

  - MedLDA Classification Model
    \[ P_2(\text{MedLDA}^\ast) : \min_{q, \alpha, \beta, \xi, \zeta} \mathcal{L}(q) + C \sum_{d=1}^{D} \xi_d \]
    s.t. \( \forall d, \forall u : E[q] \Delta D_L(u) > 1 - \xi_d, \xi_d > 0. \)
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)), LDA+SVM (baseline), MedLDA, MedLDA+SVM
- Measure: Relative Improvement Ratio

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

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

![Graph showing predictive $R^2$ and per-word log-likelihood for different models.]

Time Efficiency

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

![Graph showing CPU-Seconds for different models across different topics.]

© Eric Xing @ CMU, 2005-2013
II. Upstream Scene Understanding Models

- The “Total Scene Understanding” Model (Li et al, CVPR 2009)

Scene Classification

- 8-category sports data set (Li & Fei-Fei, 2007):
  - Fei-Fei’s theme model: 0.65 (different image representation)
  - SVM: 0.673

- 1574 images (50/50 split)
- Pre-segment each image into regions
- Region features:
  - color, texture, and location
  - patches with SIFT features
- Global features:
  - SIFT (Ding & Torralba, 2001)
  - Sparse SIFT codes (Yang et al, 2009)
MIT Indoor Scene

Classification results:
- 67-category MIT indoor scene (Quattoni & Torralba, 2009):
  - ~80 per-category for training; ~20 per-category for testing
  - Same feature representation as above
  - Gist global features

ROI+Gist(Annotation) used human annotated interest regions.

III. Supervised Multi-view MNs

A probabilistic method with an additional view of response variables $Y$

$$p(y|h) = \frac{\exp\{V^T h, y\}}{Z(V, h)}$$

Parameters can be learned with maximum likelihood estimation, e.g., special supervised Harmonium (Yang et al., 2007)
- contrastive divergence is the commonly used approximation method in learning undirected latent variable models (Welling et al., 2004; Salakhutdinov & Murray, 2008).
- t-SNE (van der Maaten & Hinton, 2008) 2D embedding of the discovered latent space representation on the TRECVID 2003 data

- Avg-KL: average pair-wise divergence

Example latent topics discovered by a 60-topic MMH on Flickr Animal Data

- Topic 1: squirrel, nature, animal, wildlife, rabbit, emo, busy, interesting
- Topic 2: wolf, ducks, animal, nature, wildlife, ducks, squirrel
- Topic 3: hawk, bird, flying, wildlife, orange, nature, finch, emo
- Topic 4: ocean, boat, animal, wildlife, diving, sea, colors, eagle, blue
- Topic 5: zebra, zoo, animal, rope, Africa, elephant, black, white, nature, emo
Classification Results

- **Data Sets:**
  - (Left) TRECVID 2003: (text + image features)
  - (Right) Flickr 13 Animal: (sift + image features)

- **Models:**
  - baseline(SVM), DWH+SVM, GM-Mixture+SVM, GM-LDA+SVM, TWH, MedLDA(sift only), MMH

Retrieval Results

- **Data Set:** TRECVID 2003
  - Each test sample is treated as a query, training samples are ranked based on the cosine similarity between a training sample and the given query
  - Similarity is computed based on the discovered latent topic representations

- **Models:** DWH, GM-Mixture, GM-LDA, TWH, MMH

- **Measure:** (Left) average precision on different topics and (Right) precision-recall curve
Infinite SVM and infinite latent SVM:

-- where SVMs meet NB for classification and feature selection

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

Mixture of SVMs

- Dirichlet process mixture of large-margin kernel machines
- Learn flexible non-linear local classifiers; potentially lead to a better control on model complexity, e.g., few unnecessary components

- The first attempt to integrate Bayesian nonparametrics, large-margin learning, and kernel methods
Infinite SVM

- RegBayes framework:

\[
\min_{p(\mathcal{M}), \xi} \text{KL}(p(\mathcal{M}) \| \pi(\mathcal{M})) - \sum_{n=1}^{N} \int \log p(x_n | \mathcal{M}) p(\mathcal{M}) d\mathcal{M} + U(\xi)
\]

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

- Model – latent class model
- Prior – Dirichlet process
- Likelihood – Gaussian likelihood
- Posterior constraints – max-margin constraints

- DP mixture of large-margin classifiers

  process of determining which classifier to use:

  1. draw \( \nu_i | \alpha \sim \text{Beta}(1, \alpha) \), \( i \in \{1, 2, \ldots\} \).
  2. draw \( \eta_i | C_0 \sim C_0 \), \( i \in \{1, 2, \ldots\} \).
  3. for the \( d \)th data point:
     a. draw \( Z_d \sim \text{Mult}(\pi(\psi)) \)

- Given a component classifier:

\[
P(y, \mathbf{x}; z, \eta) = q(z \mid \eta_i)^T f(y, \mathbf{x})
\]

- Overall discriminant function:

\[
P(y, \mathbf{x}) = \mathbb{E}_{q(z, \eta)}[P(y, \mathbf{x}; z, \eta)] = \sum_{i=1}^{\infty} q(z = i) \mathbb{E}[\eta_i] ^T f(y, \mathbf{x})
\]

- Prediction rule:

\[
y^* = \arg \max_y P(y, \mathbf{x})
\]

- Learning problem:

\[
\min_{q(\mathbf{a}, \eta)} \text{KL}(q(\mathbf{a}, \eta) \| p(\mathbf{a}, \eta)) + C_1 \mathcal{R}(q(\mathbf{a}, \eta)),
\]

\[
\mathcal{R}(q(\mathbf{a}, \eta)) = \sum_{d} \max_{\mathbf{y}, \mathbf{x}_d} \{ F(y, \mathbf{x}_d) - F(y_d, \mathbf{x}_d) \}
\]
Infinite SVM

- Assumption and relaxation
  - Truncated variational distribution
    \[ q(z, \eta, \gamma, v) = \prod_{d=1}^{D} q(z_d) \prod_{t=1}^{T} q(\eta_t) \prod_{T} q(\gamma_t) \prod_{t=1}^{T-1} q(v_t) \]
  - Upper bound the KL-regularizer

- Opt. with coordinate descent
  - For \( q(\eta) \), we solve an SVM learning problem
  - For \( q(z) \), we get the closed update rule
    \[ q(z_d = t) \propto \exp\left\{ (\beta \log v_t + \sum_{i=1}^{K} \beta \log(1 - v_i)) + \rho (\gamma_t^T x_d - \mathbb{E}[A(\gamma_t)]) + (1 - \rho) \sum_{y} \omega_{y,t} x_d^T (y) \right\} \]
    - The last term regularizes the mixing proportions to favor prediction
  - For \( q(\gamma); q(v) \), the same update rules as in (Blei & Jordan, 2006)

Experiments on high-dim real data

- Classification results and test time:
  
<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>F1 Score</th>
<th>Test Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNJL</td>
<td>49.8 ± 0.0</td>
<td>48.4 ± 0.0</td>
<td>6.02 ± 0.00</td>
</tr>
<tr>
<td>DMS JR</td>
<td>51.7 ± 0.0</td>
<td>50.1 ± 0.0</td>
<td>6.35 ± 0.01</td>
</tr>
<tr>
<td>DMS SVM</td>
<td>52.2 ± 0.0</td>
<td>48.4 ± 0.0</td>
<td>7.58 ± 0.06</td>
</tr>
<tr>
<td>DP4MNL-300</td>
<td>51.2 ± 0.0</td>
<td>49.9 ± 0.8</td>
<td>22.1 ± 7.39</td>
</tr>
<tr>
<td>DP4MNL-500</td>
<td>53.9 ± 0.0</td>
<td>53.9 ± 0.0</td>
<td>77.4 ± 20.99</td>
</tr>
<tr>
<td>DP4MNL-1000</td>
<td>53.2 ± 0.4</td>
<td>53.3 ± 0.4</td>
<td>0.22 ± 0.01</td>
</tr>
<tr>
<td>DP4MNL-1000</td>
<td>54.2 ± 0.4</td>
<td>54.6 ± 0.7</td>
<td>0.67 ± 0.05</td>
</tr>
</tbody>
</table>

  For training, linear-SVM is very efficient (~200s); RBF-iSVM is much slower, but can be significantly improved using efficient kernel methods (Rahimi & Recht, 2007; Fine & Scheinberg, 2001)

- Clusters:
  - similar background images group
  - a cluster has fewer categories
Learning Latent Features

- Infinite SVM is a Bayesian nonparametric latent class model
  - discover clustering structures
  - each data point is assigned to a single cluster/class

- Infinite Latent SVM is a Bayesian nonparametric latent feature/factor model
  - discover latent factors
  - each data point is mapped to a set (can be infinite) of latent factors
  - Latent factor analysis is a key technique in many fields; Popular models are FA, PCA, ICA, NMF, LSI, etc.

Infinite Latent SVM

- RegBayes framework:
  \[
  \min_{p(\mathcal{M}), \xi} \text{KL}(p(\mathcal{M})||\pi(\mathcal{M})) - \sum_{n=1}^{N} \int \log p(x_n|\mathcal{M})p(\mathcal{M})d\mathcal{M} + U(\xi)
  \]
  s.t.: \( p(\mathcal{M}) \in \mathcal{P}_{\text{post}}(\xi) \),

- convex function

- direct and rich constraints on posterior distribution

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

- A random finite binary latent feature model
  \[ \pi_k | \alpha \sim \text{Beta}\left(\frac{\alpha}{K}, 1\right) \]
  \[ z_{ik} | \pi_k \sim \text{Bernoulli}(\pi_k) \]

  - \( \pi_k \) is the relative probability of each feature being on, e.g.,

  - \( z_{it} \), are binary vectors, giving the latent structure that's used to generate the data, e.g.,
  \[ x_t \sim \mathcal{N}(\eta^T z_{it}, \delta^2) \]

Indian Buffet Process

- A stochastic process on infinite binary feature matrices

  Generative procedure:
  - Customer 1 chooses the first \( K_1 \) dishes: \( K_1 \sim \text{Poisson}(\alpha) \)
  - Customer \( i \) chooses:
    - Each of the existing dishes with probability \( \frac{m_k}{i} \)
    - \( K_i \) additional dishes, where \( K_i \sim \text{Poisson}\left(\frac{\alpha}{i}\right) \)

  \( Z_{it} \sim \text{IBP}(\alpha) \)
Posterior Constraints – classification

- Suppose latent features $z$ are given, we define latent discriminant function:
  \[ f(y, x, z; \eta) = \eta^T g(y, x, z) \]

- Define effective discriminant function (reduce uncertainty):
  \[ f(y, x; p(Z, \eta)) = \mathbb{E}_{p(Z, \eta)}[f(y, x, z; \eta)] = \mathbb{E}_{p(Z, \eta)}[\eta^T g(y, x, z)] \]

- Posterior constraints with max-margin principle
  \[ \forall n \in \mathcal{N}_tr, \forall y : \quad f(y_n, x_n; p(Z, \eta)) - f(y, x_n; p(Z, \eta)) \geq \ell(y, y_n) - \zeta_n \]

Experimental Results

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

<table>
<thead>
<tr>
<th>Model</th>
<th>TRECVID2003</th>
<th>Flickr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>F1 score</td>
</tr>
<tr>
<td>ELF-NVM</td>
<td>0.565 ± 0.0</td>
<td>0.412 ± 0.0</td>
</tr>
<tr>
<td>MLH</td>
<td>0.566 ± 0.0</td>
<td>0.150 ± 0.0</td>
</tr>
<tr>
<td>IHP-NVM</td>
<td>0.553 ± 0.013</td>
<td>0.597 ± 0.000</td>
</tr>
<tr>
<td>uLSVM</td>
<td>0.563 ± 0.010</td>
<td>0.418 ± 0.011</td>
</tr>
</tbody>
</table>
Summary

Bayesian kernel machines; Infinite GPs

Large-margin learning

Large-margin kernel machines

Infinite SVM (iSVM)  Infinite Latent SVM (iLSVM)

Bayesian estimation

Kernel methods

Linear Expectation Operator (resolve uncertainty)

Large-margin learning
Summary

- A general framework of MaxEnDNet for learning structured input/output models
  - Subsumes the standard M3Ns
  - Model averaging: PAC-Bayes theoretical error bound
  - Entropic regularization: sparse M3Ns
  - Generative + discriminative: latent variables, semi-supervised learning on partially labeled data, fast inference

- PoMEN
  - Provides an elegant approach to incorporate latent variables and structures under max-margin framework
  - Enable Learning arbitrary graphical models discriminatively

- Predictive Latent Subspace Learning
  - MedLDA for text topic learning
  - Med total scene model for image understanding
  - Med latent MNs for multi-view inference

- Bayesian nonparametrics meets max-margin learning

- Experimental results show the advantages of max-margin learning over likelihood methods in EVERY case.

Remember: Elements of Learning

- Here are some important elements to consider before you start:
  - Task:
    - Embedding? Classification? Clustering? Topic extraction? …
  - Data and other info:
    - Input and output (e.g., continuous, binary, counts, …)
    - Supervised or unsupervised, of a blend of everything?
    - Prior knowledge? Bias?
  - Models and paradigms:
    - BN? MRF? Regression? SVM?
    - Bayesian/Frequents? Parametric/Nonparametric?
  - Objective/Loss function:
    - MLE? MCLE? Max margin?
    - Log loss, hinge loss, square loss? …
  - Tractability and exactness trade off:
    - Online? Batch? Distributed?
  - Evaluation:
    - Visualization? Human interpretability? Perplexity? Predictive accuracy?

- It is better to consider one element at a time!