Dictionary Learning

Jun Zhu
dcszj@mail.tsinghua.edu.cn
State Key Lab of Intelligent Technology & Systems
Tsinghua University

December 20, 2012
Today’s Plan

- Review Sparse Learning

- Dictionary Learning
  - The basic settings and algorithms

- Applications
  - Images/Videos
  - Text Documents
Two Important Aspects

- Model goodness:
  - Often defined in terms of prediction accuracy
- Model parsimony:
  - Simpler models are preferred for the sake of scientific insight into the $x - y$ relationship
Supervised Learning and Regularization

**Data**

\[ x_i \in \mathcal{X}, \ y_i \in \mathcal{Y}, \ i = 1, 2, \ldots, n \]

Minimize with respect to function \( f : \mathcal{X} \rightarrow \mathcal{Y} \)

\[
\sum_{n} \ell(f(x_i), y_i) + \frac{\lambda}{2} \| f \|^2
\]

Error on data + Regularization

**Two theoretical/algorithmic issues**

- Loss
- Function space/norm
Sparse Linear Estimation with $\ell_1$-norm

The general setting $f : \mathcal{X} \rightarrow \mathcal{Y}$

$$\sum_{n} \ell(f(x_i), y_i) + \Omega(f)$$

Sparse linear estimation with $\ell_1$-norm

$$f(x) = w^\top x$$

$$\Omega(f) = \|w\|_1 = \sum_{i=1}^{p} |w_i|$$
Why $\ell_1$-norm leads to sparsity?

**Example 1**: quadratic problem in 1D

$$\min_x \frac{1}{2} x^2 - xy + \lambda |x|$$

- Piecewise quadratic function with a kink at zero
  - Derivative at 0+: $g_+ = -y + \lambda$
  - Derivative at 0-: $g_- = -y - \lambda$
  - $x = 0$ is the solution iff $g_+ \geq 0$, $g_- \leq 0$ (i.e.: $|y| \leq \lambda$)
  - $x \geq 0$ is the solution iff $g_+ \leq 0$ (i.e.: $y \geq \lambda$) $x^* = y - \lambda$
  - $x \leq 0$ is the solution iff $g_- \geq 0$ (i.e.: $y \leq -\lambda$) $x^* = y + \lambda$

- Solution is: $x^* = \text{sign}(y)(|y| - \lambda)_+$ | Soft Thresholding
Why $\ell_1$-norm leads to sparsity?

- **Example 1**: quadratic problem in 1D

  $$\min_x \frac{1}{2}x^2 - xy + \lambda|x|$$

- Piecewise quadratic function with a kink at zero

- Solution is: **Soft Thresholding**

  $$x^* = \text{sign}(y)(|y| - \lambda)_+$$
Why $\ell_1$-norm leads to sparsity?

**Example 2:** minimize quadratic function $Q(w)$ subject to 

$$\|w\|_1 \leq T$$

**Geometric Interpretation**

- Penalizing is “equivalent” to constraining
Revisit the Diabetes Study

Lasso: least absolute shrinkage and selection operator

\[
\min_w L(y, w^\top X) = \sum_i (y_i - w^\top x_i)^2 \\
\text{s.t.: } \|w\|_1 \leq t
\]
Nonsmooth convex analysis & optimization

- Analysis
  - optimal conditions
- Optimization
  - algorithms
Optimal conditions for smooth optimization – **Zero Gradient**

**Example:**

\[
\min_w J(w) = \sum_{i=1}^n \ell(y_i, w^\top x_i) + \frac{\lambda}{2} \|w\|_2^2
\]

- gradient

\[
\nabla J(w) = \sum_i \nabla_w \ell(y_i, w^\top x_i) x_i + \lambda w
\]

- If squared loss

\[
\sum_i \ell(y_i, w^\top x_i) = \frac{1}{2} \|y - Xw\|_2^2
\]

  - gradient

\[
\nabla J(w) = -X^\top (y - Xw) + \lambda w
\]

  - solution

\[
w = (\lambda I + X^\top X)^{-1} X^\top y
\]

**But \(l_1\)-norm is non-differentiable**

- Can’t compute the gradient \(\Rightarrow\) subgradient (directional derivatives)
Directional Derivatives

- Directional derivative in direction $\Delta$ at $w$:

$$\nabla J(w, \Delta) = \lim_{\epsilon \to 0^+} \frac{J(w + \epsilon \Delta) - J(w)}{\epsilon}$$

- Rate of change moving through $w$ at the velocity specified by $\Delta$
- Always exist when $J$ is convex and continuous

**Main idea:** in non-smooth settings, may need to look at all directions

- **Proposition:** $J$ is differentiable at $w$ iff $\Delta \mapsto \nabla J(w, \Delta)$ is linear

$$\nabla J(w, \Delta) = \nabla J(w)^\top \Delta$$
Optimal conditions for convex functions

- Unconstrained minimization
  - Proposition: \( w \) is optimal iff
    \[
    \forall \Delta \in \mathbb{R}^p : \nabla J(w, \Delta) \geq 0
    \]
  - i.e., function value goes up in all directions

- Reduces to zero-gradient for smooth problems
Directional derivative for $\ell_1$-norm

- **Function**

$$J(w) = \sum_i \ell(y_i, w^\top x_i) + \lambda \|w\|_1 = L(y, Xw) + \lambda \|w\|_1$$

- **$\ell_1$-norm:**

$$\|w + \epsilon \Delta\|_1 - \|w\|_1 = \sum_{j, w_j \neq 0} (|w_j + \epsilon \Delta_j| - |w_j|) + \sum_{j, w_j = 0} |\epsilon \Delta_j|$$

- Thus (separability of optimal conditions)

$$\nabla J(w, \Delta) = \nabla L(w)^\top \Delta + \lambda \sum_{j, w_j \neq 0} \text{sign}(w_j) \Delta_j + \lambda \sum_{j, w_j = 0} |\Delta_j|$$

$$= \sum_{j, w_j \neq 0} (\nabla L(w)_j + \lambda \text{sign}(w_j)) \Delta_j + \sum_{j, w_j = 0} (\nabla L(w)_j \Delta_j + \lambda |\Delta_j|)$$
Directional derivative for $\ell_1$-norm

- **General loss**: $w$ is optimal if and only if for all $j = 1, 2, \ldots, p$

  $$w_j \neq 0 \Rightarrow \nabla L(w)_j + \lambda \text{sign}(w_j) = 0$$

  $$w_j = 0 \Rightarrow |\nabla L(w)_j| \leq \lambda$$

- **Squared loss**: $L(y, Xw) = \sum_i (y_i - w^\top x_i)^2$

  $$\nabla L(w)_j = -X_j^\top (y - Xw)$$

- $X_j$ is the $j$-th column of $X$
First-order methods for convex optimization – smooth optimization

- **Gradient descent:**

  \[ w_{t+1} = w_t - \alpha_t \nabla J(w_t) \]

  - with line search: search for a descent \( \alpha_t \)
  - with fixed step size, e.g., \( \alpha_t = a(t + b)^{-1} \)

- **Convergence of** \( f(w_t) \) **to** \( f^*(w) = \min_w f(w) \)

  - Depends on the condition number of the optimization number (i.e., correlation within variables)

- **Coordinate descent:**

  - Similar properties
Regularized problems – proximal methods

- Gradient descent as a proximal method

\[ w_{t+1} = \arg \max_w L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{\mu}{2} \|w - w_t\|^2_2 \]

\[ = w_t - \frac{1}{\mu} \nabla L(w_t) \]

- Regularized problems of the form \( \min_w L(w) + \lambda \Omega(w) \)

\[ w_{t+1} = \arg \max_w L(w_t) + (w - w_t)^\top \nabla L(w_t) + \lambda \Omega(w_t) + \frac{\mu}{2} \|w - w_t\|^2_2 \]

\[ = \text{SoftThreshold}(w_t - \frac{1}{\mu} \nabla L(w_t)) \]

- Similar convergence rates as smooth optimization
  - Acceleration methods (Nestrov, 2007; Beck & Teboulle, 2009)
**More on Proximal Mapping**

The **proximal mapping** (or proximal operator) of a convex function $h$ is

$$\text{prox}_h(x) = \arg\min_{\mu} (h(\mu) + \frac{1}{2}\|\mu - x\|^2_2)$$

**Examples:**

- $h(x) = 0$ : $\text{prox}_h(x) = x$
- $h(x) = I_C(x)$ (indicator function of $C$): a projection on $C$
  $$\text{prox}_h(x) = P_C(x) = \arg\min_{\mu \in C} \|\mu - x\|^2_2$$
- $h(x) = t\|x\|_1$: a shrinkage (soft-threshold) operation
  $$\text{prox}_h(x)_i = \begin{cases} 
  x_i - t & x_i \geq t \\
  0 & |x_i| \leq t \\
  x_i + t & x_i \leq -t 
\end{cases}$$
More on Proximal Gradient Methods

- **Unconstrained problem** with cost function split in two parts
  \[ \min_x f(x) = g(x) + h(x) \]
  - \( g \) is convex, differentiable
  - \( h \) closed, convex, possibly nondifferentiable; \( \text{prox}_h \) is inexpensive

- **Proximal gradient algorithm:**
  \[ x^{(k+1)} = \text{prox}_{t_k h} \left( x^{(k)} - t_k \nabla g(x^{(k)}) \right) \]
  - \( t_k \) is step size, constant or determined by line search
More on Proximal Gradient Methods

From definition of proximal operator

\[ x^{(k+1)} = \text{prox}_{t_k h} \left( x^{(k)} - t_k \nabla g(x^{(k)}) \right) \]

\[
\begin{align*}
    x^{(k+1)} &= \arg \min_{\mu} \left( h(\mu) + \frac{1}{2t_k} \| \mu - x^{(k)} + t_k \nabla g(x^{(k)}) \|_2^2 \right) \\
    &= \arg \min_{\mu} \left( h(\mu) + g(x^{(k)}) + \nabla g(x^{(k)})^\top (\mu - x^{(k)}) + \frac{1}{2t_k} \| \mu - x^{(k)} \|_2^2 \right)
\end{align*}
\]

- i.e., minimizes \( h(\mu) \) plus a simple quadratic local model of \( g(\mu) \) around \( x^{(k)} \)
Comparison on Algorithms for Lasso

\[ n = 2000, p = 10,000 \]

(a) corr: low, reg: low

- SG: sub-gradient descent
- Ista: simple proximal methods
- Fista: accelerated version of Ista
- Re-L2: reweighted least square
- CD: coordinate descent

(b) corr: low, reg: high

- CP: cone programming
- QP: quadratic programming
- Lars: least angle regression
- CD: coordinate descent
Alternative sparse methods

– Bayesian methods

- Heavy-tailed priors as a proxy to sparsity
  - Student distributions (Caron and Doucet, 2008)
  - Generalized hyperbolic priors (Archambeau and Bach, 2008)
  - Instance of automatic relevance determination (Neal, 1996)

- Spike and Slab (Ishwaran and Rao, 2005)
  - Mixtures of “Diracs” and another absolutely continuous distributions

\[ p(\sigma^2) = \alpha \delta(\sigma^2 = \nu_0) + (1 - \alpha)\text{Gamma}(a, b) \]

where \( \nu_0 \) is a small near zero value

- Less theory than frequentist methods
Regularization with Groups of Variables

\[ \Omega(w) = \sum_{i=1}^{m} \| w_{G_i} \|_2 \]

E.g.: \[ \| (w_1, w_2) \|_2 + \| w_3 \|_2 \leq 1 \]
Group Lasso

- Opt. problem:
  \[
  \min_w \sum_i (y_i - w^\top x_i)^2 + \lambda \sum_{i=1}^m \sqrt{p_i} \|w_{G_i}\|_2
  \]

- Optimal condition:
  - Proposition: \( w \) is optimal iff \( \forall j = 1, 2, \ldots, m \)
    \[
    w_{G_j} \neq 0 \Rightarrow -X_{G_j}^\top (y - Xw) + \frac{\lambda \sqrt{p_j} w_{G_j}}{\|w_{G_j}\|_2} = 0
    \]
    \[
    w_{G_j} = 0 \Rightarrow \|X_{G_j}^\top (y - Xw)\|_2 \leq \lambda \sqrt{p_j}
    \]
  - \( p_j \) is the number of features in group \( j \).
  - Coordinate descent algorithm can be used to solve it.
Two types of sparsity of matrices

- **Type 1 of sparsity:**
  - Directly on the elements

Many elements are zeros \hspace{2cm} Many rows or columns are zeros
Two types of sparsity of matrices

**Type 2 of sparsity:**
- Through a factorization $M = UV^\top \in \mathbb{R}^{n \times m}$, $V \in \mathbb{R}^{p \times m}$
- Low-rank sparsity: $m$ is small
- Sparse decomposition: $U$ sparse

\[ M = U \quad \text{and} \quad V^\top \]

\[ M = U \quad \text{and} \quad V^\top \]
Today’s Plan

- Review Sparse Learning

- Dictionary Learning
  - The basic settings and algorithms

- Applications to Text Documents and Images
Learning with pre-defined basis functions -- generalized linear models

- A mapping function
  \[ \phi : \mathcal{X} \rightarrow \mathbb{R}^N \]

- Doing linear regression in the mapped space
  \[ f(x) = \phi(x)^\top w \]

\[ h: x \rightarrow h(x) \]
Fixed Basis Functions

Given a set of basis functions \( \{\phi_h(x)\}_{h=1}^H \)

\[
\phi(x) = [\phi_1(x) \cdots \phi_H(x)]^\top
\]

- E.g. 1:
  \[
  \phi_h(x) = \exp \left( - \frac{||x - c_h||^2_2}{2r^2} \right)
  \]

- E.g. 2:
  \[
  \phi_h(x) = x_i^p x_j^q
  \]

\[
f(x) = \phi(x)^\top w
\]
Dictionary Learning

Goal:
- learn the basis functions from data
Parametric Basis Functions

- Neural networks to learn a parameterized mapping function
- E.g., a two-layer feedforward neural networks

\[ \phi_h(x) = \tanh\left( \sum_{i=1}^{I} w_{hi}^{(1)} x_i + w_{h0}^{(1)} \right) \]

\[ f(x; w) = \sum_{h=1}^{H} w_{h}^{(2)} \phi_h(x) + w_0^{(2)} \]

[Figure by Neal]
PCA: minimum error formulation

- A set of complete orthonormal basis
  \[ \{\mu_i\}, \ i = 1, \ldots, D \]
  \[ \mu_i^\top \mu_j = \delta_{ij} \]

- We consider a low-dimensional approximation
  \[ \tilde{x}_n = \sum_{i=1}^{M} z_{ni} \mu_i + \sum_{i=M+1}^{D} b_i \mu_i \]

- The best approximation is to minimize the error
  \[ J = \frac{1}{N} \sum_{n=1}^{N} \|x_n - \tilde{x}_n\|^2 \]
Issues with PCA

Principal components calculated on 8x8 image patches

- PCA capture linear pairwise statistics
- Suitable for Gaussian distributed data
- Not localized
- Not resemble cortical receptive fields
- Not suitable for images with high order statistics

[Olshausen & Field, Nature 1996]
Sparse Coding

- **Basic assumption 1**: a linear superposition model
  \[ I(x, y) = \sum_i \alpha_i \phi_i(x, y) \]
  - an image
  - basis function

- **Basic assumption 2**: nature images have ‘sparse structure’
  (similar as minimum-entropy code)

\[
\alpha = \begin{pmatrix}
0.4 \\
0 \\
0 \\
0.1 \\
0.2 \\
0 \\
\vdots \\
0
\end{pmatrix}
\]

[Olshausen & Field, Nature 1996]
Sparse Coding

Search for a sparse code is an optimization problem:

$$\min_{\alpha, \phi} \sum_{x, y} \left[ I(x, y) - \sum_i \alpha_i \phi_i(x, y) \right]^2 + \psi(\alpha)$$

- the L1-norm is a common choice

Solve the problems — alternating minimization

- For each image, solve for $\alpha$ as a sparse learning problem
- Update dictionary using gradient descent

[Olshausen & Field, Nature 1996]
Sparse Coding

- Basis learned on 16 x 16 natural scene image patches
  
  - Localized
  - Oriented
  - Selective to spatial scales

[Olshausen & Field, Nature 1996]
Nonnegative Matrix Factorization

Matrix factorization

$$\min_{U \in \mathbb{R}^{n \times m}, V \in \mathbb{R}^{p \times m}} L(UV^T; X)$$

Example losses:

$$L = \sum_{i=1}^{n} \sum_{j=1}^{p} ((UV^T)_{ij} - X_{ij})^2$$

$$L = \sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} \log(UV^T)_{ij} - (UV^T)_{ij})$$

Non-negativity loss:

$$U \geq 0; \ V \geq 0$$

[Lee & Seung, Nature 1999]
Nonnegative Matrix Factorization

- Sparse basis and sparse coefficients for images:

\[ \text{Original} \times \begin{bmatrix} \text{NMF} \\ \end{bmatrix} = \]

[Lee & Seung, Nature 1999]
Nonnegative Matrix Factorization

Eigenfaces and non-sparse coefficients by PCA
- Positive and negative combinations

[Lee & Seung, Nature 1999]
Nonnegative Matrix Factorization

- NMF for text documents with bag-of-word counts

\[ X \approx UV \]

\[ X = \begin{pmatrix}
    x_{11} & x_{12} & \cdots & x_{1D} \\
    x_{21} & x_{22} & \cdots & x_{2D} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{N1} & x_{N2} & \cdots & x_{ND}
\end{pmatrix}_{N \times D} \]

\[ X_d = \begin{pmatrix}
    X_{1d} \\
    X_{2d} \\
    \vdots \\
    X_{Nd}
\end{pmatrix}_{N \times 1} \approx \begin{pmatrix}
    \vdots \\
    U_{.1} & U_{.2} & \cdots & U_{.K} \\
    \vdots
\end{pmatrix}_{N \times K} \times \begin{pmatrix}
    V_{1d} \\
    V_{2d} \\
    \vdots \\
    V_{Kd}
\end{pmatrix}_{K \times 1} \]

- The same coefficient vector to reconstruct all word counts in a document
Nonnegative Matrix Factorization

- NMF for text documents with bag-of-word counts

<table>
<thead>
<tr>
<th>court</th>
<th>president</th>
</tr>
</thead>
<tbody>
<tr>
<td>government</td>
<td>served</td>
</tr>
<tr>
<td>council</td>
<td>governor</td>
</tr>
<tr>
<td>culture</td>
<td>secretary</td>
</tr>
<tr>
<td>supreme</td>
<td>senate</td>
</tr>
<tr>
<td>constitutional</td>
<td>congress</td>
</tr>
<tr>
<td>rights</td>
<td>presidential</td>
</tr>
<tr>
<td>justice</td>
<td>elected</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>flowers</th>
<th>disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>leaves</td>
<td>behaviour</td>
</tr>
<tr>
<td>plant</td>
<td>glands</td>
</tr>
<tr>
<td>perennial</td>
<td>contact</td>
</tr>
<tr>
<td>flower</td>
<td>symptoms</td>
</tr>
<tr>
<td>plants</td>
<td>skin</td>
</tr>
<tr>
<td>growing</td>
<td>pain</td>
</tr>
<tr>
<td>annual</td>
<td>infection</td>
</tr>
</tbody>
</table>

Encyclopedia entry: 'Constitution of the United States'

- president (148)
- congress (124)
- power (120)
- united (104)
- constitution (81)
- amendment (71)
- government (57)
- law (49)

[Lee & Seung, Nature 1999]
Today’s Plan

- Review Sparse Learning

- Dictionary Learning
  - The basic settings and algorithms

- Applications to Text Documents and Images
Topic Modeling – projection view

Dictionary Learning

The room was \textit{dirty} and smelled \textit{awful}. The wallpaper was \textit{dirty} and \textit{nasty}. The hallway smelled \textit{awful} as did the water. The price was \textit{awful} 100.00 to stay in such a \textit{dirty} room. Just \textit{awful}. I would not recommend this hotel to an \textit{enemy}.

\textbf{Fantastic} Hotel in \textit{beautiful} Krabi. We felt pampered by the \textit{gentle} and \textit{caring} staff. The turndown service each evening maintains the very \textit{high} standard. A \textit{blissful} sleep in the most \textit{comfortable} beds. The food is of a extremely \textit{high} quality with \textit{fantastic} \textit{fresh} and \textit{vibrant} tastes.

Topical Projection

\textbf{Probabilistic topic models}:
- Topical subspace is a simplex
- Projection under KL-divergence

\textbf{spanned topical subspace}
Probabilistic Topic Models – restrictions

- Ineffective in controlling posterior sparsity by using priors, e.g., Dirichlet prior in LDA (Zhu & Xing, UAI 2011):

  - Restricted to MLE when considering supervised side information;
  - Hard in inference due to a normalized likelihood model when considering discrete side information (e.g., category labels or features)

  ![Graph showing weak and strong smoothing]

  **weak smoothing:** sparse but worse predict accuracy
  **strong smoothing:** improved prediction, but non-sparse
Topical Modeling with Rich Features

Bag-of-Words

My wife and I spent a week at this Fantastic Unsurpassed Hotel in beautiful Krabi. From the moment we arrived to the morning of our departure we felt pampered by the gentle and caring staff that look after your of a quality that other hotels could only dream of, the turndown service each eve small deserts or petit fours for your taste before a blissful sleep in the most restaurants is of a extremely high quality with fantastic fresh and vibrant taste hotel make you realize just how lucky you are to be in such paradise. Thank you all at The Amari Vogue Resort.

Spent the night at the xxxx last night. The room was dirty and smelled awful. I ripping out the baseboards, tore the wallpaper which was dirty and nasty. To brush my teeth. We spent 10 days in Cambodia—a 3rd world country—in Stung treng was awful 100.00 to stay in such a dirty room. Just awful, I would not recom
Topical Modeling with Rich Features

Bag-of-Words + Features (POS tag, WordNet, etc)

My wife and I spent a week at this Fantastic Unsurpassed Hotel in beautiful departure we felt pampered by the gentle and caring staff that look after you of a quality that other hotels could only dream of, the turndown service each small deserts or petit fours left for your taste before a blissful sleep in the most restaurants is of a extremely high quality with fantastic fresh and vibrant hotel make you realize just how lucky you are to be in such paradise. Thank

Spent the night at the xxxx last night. The room was dirty and smelled awful ripping out the baseboards, tore the wallpaper which was dry and nasty. Told to brush my teeth. We spent 10 days in Cambodia--a 3rd world country---in St was awful 100.00 to stay in such a dirty room. Just awful, I would not reco
Probabilistic Conditional Topic Models

Logistic-Normal Topic Models with Features

\[ \theta_d \sim \mathcal{N}(\mu, \Sigma) \]

\[ p(z_m|\theta, a) = \frac{1}{B(\theta)} \exp \left( \theta^\top f(z_m, a) \right), \text{ where } B(\theta) = \sum_z \exp(\theta^\top f(z, a)) \]

- Problem of non-conjugacy:
  - sum-exp function can make the inference & training hard
  - variational approximation can be expensive

Can we avoid normalization?
**Sparse Topical Coding (STC)**

**A Non-probabilistic Topic Model**

- **Topical bases:**
  \[ \beta_k \in \mathcal{P}_N, \quad \beta = \begin{pmatrix} \beta_{1} & \beta_{2} & \cdots & \beta_{N} \end{pmatrix}_{K \times N} \]

- **Hierarchical coding:**
  - **word code** \( s \) – encode word counts under a loss:
    \[ \ell(w_n, s_n, \beta) = \log p(w_n|s_n, \beta) \]
    where \( \mathbb{E}_{p(w_n|s_n, \beta)}[T(w_n)] = s_n^\top \beta_n \) – we use *Poisson* distribution
  - **document code** \( \theta \) – an aggregation of word codes
    \[ \sum_{n \in I} \| s_n - \theta \|_2^2 \]

- **Nonnegative hierarchical sparse coding (with dictionary learning)**
  \[
  \min_{\{s_{dn}, \theta_d\}, \beta} \sum_{d, n \in I_d} \ell(w_{dn}, s_{dn}^\top \beta_n) + \sum_{d, n \in I_d} \left( \frac{\gamma}{2} \| s_{dn} - \theta_d \|_2^2 + \rho \| s_{dn} \|_1 \right) + \lambda \sum_d \| \theta_d \|_1
  \]
  subject to: \( \beta_k \in \mathcal{P}_N, \forall k; \theta_d \geq 0, \forall d; s_{dn} \geq 0, \forall d, n \in I_d; \)
Sparse Topical Coding (STC)

A Projection View

(unnormalized) KL-divergence for log-Poisson loss

Projection is done under Regularization!
Conditional Topical Coding (CTC)

Conditional features

- Dual-view model for word codes

\[
\text{STC: } \frac{\gamma}{2} \| \mathbf{s}_n - \theta \|^2_2 + \rho \| \mathbf{s}_n \|_1
\]

\[
\text{CTC: } \frac{\gamma}{2} (\| \mathbf{s}_n - \theta \|^2_2 + \| \mathbf{s}_n - \mathbf{U} \mathbf{f}(\mathbf{a}) \|^2_2) + \rho \| \mathbf{s}_n \|_1
\]

\[
\mathbf{f}(\mathbf{a}) = \begin{pmatrix}
\mathbf{f}_1(\mathbf{a}) \\
\mathbf{f}_2(\mathbf{a}) \\
\vdots \\
\mathbf{f}_L(\mathbf{a})
\end{pmatrix}, \text{ where } f_l : \mathcal{A} \to \mathbb{R} \quad \mathbf{U} = \begin{pmatrix}
\mathbf{U}_1 \\
\mathbf{U}_2 \\
\vdots \\
\mathbf{U}_L
\end{pmatrix}_{K \times L}
\]
Max-margin Conditional Topical Coding

- **Supervised side information**
  - Widely available side information
    - review rating scores on TripAdvisor
    - image categories in LabelMe, ImageNet
  - Regression:
    \[ y \in \mathbb{R} \]
  - Non-probabilistic STC/CTC seamlessly incorporate any convex loss function, which may not arise from a probabilistic model

\[
\min_{\Theta, U, \eta} \ h(\Theta, U) + C\mathcal{R}(\{\theta_d\}, \eta) + \frac{1}{2} \|\eta\|_2^2
\]

s.t.: constraints

where \( \mathcal{R}_D = \frac{1}{D} \sum_d \max(0, |\eta^\top \theta_d - y_d| - \epsilon) \) is \( \epsilon \)-insensitive loss for support vector regression with latent document codes \( \{\theta_d\} \)
Coordinate Descent Algorithm

- Not jointly convex, but convex over subsets of parameters
- Alternately solve four sub-problems

convex for each sub-problem
Step 1: hierarchical sparse coding – words

We solve for the uncoupled word codes:

$$\min_{s_n \geq 0} \ell(w_n, s_n^T \beta_n) + \frac{\gamma}{2} (\|s_n - \theta\|_2^2 + \|s_n - Uf(a)\|_2^2) + \rho \|s_n\|_1$$

**Proposition 1:**

Let $\psi(x)$ be a convex function on $\mathbb{R}$. If $x_0$ is a solution of the unconstrained problem $P_0 : \min_x \psi(x)$, then $x^* = \max(0, x_0)$ is an optimal solution of the constrained problem $P_1 : \min_{x \geq 0} \psi(x)$.

$$P_0 : \min_{s_n} \ell(w_n, s_n^T \beta_n) + \frac{\gamma}{2} (\|s_n - \theta\|_2^2 + \|s_n - Uf(a)\|_2^2) + \rho \sum_k s_{nk}$$

We take the larger real solution
Step 1: hierarchical sparse coding – docs

We solve for each document code:

\[
\min_{\theta \geq 0} \lambda \| \theta \|_1 + \frac{\gamma}{2} \sum_{n \in I} \| s_n - \theta \|_2^2 + \frac{C}{D} \max(0, |\eta^\top \theta - y| - \epsilon)
\]

- By proposition 1, one solution using sub-gradient:

\[\forall k, \quad \theta_k = \max(0, \bar{s}_k - \frac{\lambda}{\gamma|I|})\]

where \(\bar{s}_k = \frac{1}{|I|} \sum_{n \in I} s_{nk} - \frac{C}{D|I|} \mathbb{I}(|\eta^\top \theta - y| > \epsilon) \text{Sign}(\eta^\top \theta - y)\eta_k\)
Step 2: prediction model learning

$$\min_{\eta} \quad CR_D(\{\theta_d\}, \eta) + \frac{1}{2} \|\eta\|_2^2$$

where

$$R_D = \frac{1}{D} \sum_d \max(0, |\eta^\top \theta_d - y_d| - \epsilon)$$

$$\min_{\eta, \xi, \xi^*} \quad C \frac{1}{D} \sum_d (\xi_d + \xi_d^*) + \frac{1}{2} \|\eta\|_2^2$$

s.t.: $\begin{cases} y_d - \eta^\top \theta_d & \leq \epsilon + \xi_d \\ -y_d + \eta^\top \theta_d & \leq \epsilon + \xi_d^* \end{cases}$

- Support vector regression with learned latent features $\theta$
- SVMLight is a high-performance package for solving this problem
Step 3: feature weight learning

\[
\min_{\mathbf{U}} \sum_{d,n \in I_d} \frac{\gamma}{2} \| \mathbf{s}_{dn} - \mathbf{U} f(\mathbf{a}_d) \|_2^2 + \frac{1}{2} \| \mathbf{U} \|_2^2
\]

\[
\begin{pmatrix}
  f_1(\mathbf{a}) \\
  f_2(\mathbf{a}) \\
  \vdots \\
  f_L(\mathbf{a})
\end{pmatrix}, \text{ where } f_l : \mathcal{A} \to \mathbb{R} \quad \mathbf{U} = \begin{pmatrix}
  \mathbf{U}_{.1} & \mathbf{U}_{.2} & \cdots & \mathbf{U}_{.L}
\end{pmatrix}_{K \times L}
\]

- ridge regression for each feature:

\[
\min_{\mathbf{U}_l} \sum_{d,n \in I_d} \frac{\gamma}{2} \| \mathbf{s}_{dn} - f_l(\mathbf{a}_d) \mathbf{U}_l \|_2^2 + \frac{1}{2} \| \mathbf{U}_l \|_2^2
\]

- analytical solution or efficiently solved with gradient descent
Step 4: dictionary learning

\[
\min_{\beta} \sum_{d,n\in I_d} \ell(w_n, s_{d_n}^{\top} \beta_n)
\]

s.t. : \(\beta_k \in \mathcal{P}_N, \forall k\)

- convex objective with linear constraints
- projected gradient descent, where the projection to simplex can be done linearly (Duchi et al., 2008)
Coordinate Descent Algorithm

- Empirically, converges nicely to local optimum
Experiments:
Sparse Topical Coding

Data Sets:
- 20 Newsgroups
- Documents from 20 categories
- ~20,000 documents in each group
- Remove stop word as listed in UMASS Mallet
Prediction Accuracy on 20Newsgroups

- gaussSTC: uses L2-norm regularizer on word and doc codes
- NMF: non-negative matrix factorization
- regLDA: LDA model using entropic regularizer on topic assignment distributions
- MedLDA: max-margin supervised LDA (Zhu et al., 2012)
- DiscLDA: discriminative LDA (Simon et al., 2008)
Sparsity of Word Codes on 20Newsgroups

- Sparsity ratio: the percentage of zero elements on the word codes
## Sparse Word Codes on 20Newsgroups

<table>
<thead>
<tr>
<th>Class</th>
<th>STC</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>alt. atheism</td>
<td><img src="image" alt="god" /></td>
<td><img src="image" alt="god" /></td>
</tr>
<tr>
<td>comp. graphics</td>
<td><img src="image" alt="image" /></td>
<td><img src="image" alt="image" /></td>
</tr>
<tr>
<td>rec. autos</td>
<td><img src="image" alt="car" /></td>
<td><img src="image" alt="car" /></td>
</tr>
<tr>
<td>sport. baseball</td>
<td><img src="image" alt="baseball" /></td>
<td><img src="image" alt="baseball" /></td>
</tr>
</tbody>
</table>
Train-time on 20Newsgroups
Experiments:
Conditional Topical Coding

- TripAdvisor review (rating from 1 – 5)
  - Small dataset – 5,000 documents
  - Large dataset – 97,948 documents

- Vocabulary:
  - 12,000 terms

- Text features (Binary):
  - POS tagging:
    - adjective, noun, adverb and verb
  - WordNet:
    - positive or negative words with initial seeds based on Synonym and Antonym relationship
    - Denying word (e.g., not, never) appears in front?
  - NP chunking:
    - pair-wise constraints

---

**Lovely welcoming staff, good rooms that give a good nights sleep, downtown location**

**Meramees Hostel**

SheikhSahib ☿ 10 contributions
London
Jul 7, 2009 | Trip type: Friends getaway

This hotel is just of the side streets of Talat Herb, 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

**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

downloaded in Dec, 2009
Experiments – methods

Supervised topic models for regression

- Regression problem (take the logarithm transformation)

- Feature-based models – CTC (POSTag + WordNet):
  - CdTM: conditional topic models with max-margin training (Zhu & Xing, 2010) (POSTag + WordNet)
  - CTRF: conditional topic random fields with max-margin training (Zhu & Xing, 2010) (POSTag + WordNet + NP Chunking)

- BOW-based models – STC:
  - sLDA: supervised topic model with MLE learning (Blei & McAuliffe, 2007)
Results on Small Dataset

Predictive $R^2$: \[ pR^2 = 1 - \frac{MSE}{Data \ Variance} \]

- **Feature-based models:**
  - CTC is comparable to CTRF on best performance, better than CdTM;
  - CTC is more robust against the change of topic numbers (difficult approximate inference in CTRF)

- **BOW-based models:**
  - STC is better than MedLDA and sLDA (regularity and sparsity)
Results on Small Dataset

Test (inference) time:

- Feature-based models:
  - CTC is \(~100\) times faster than CTRF and CdTM
- BOW-based models:
  - STC is \(~10\) times faster than MedLDA and sLDA

2,500 documents
Results on Small Dataset

- Training (both inference & dictionary learning) time:

  2,500 documents

- Feature-based models:
  - CTC is \(~10\) times faster than CTRF and CdTM

- BOW-based models:
  - STC is \(~10\) times faster than MedLDA and sLDA, when the topic number is large
Results on Small Dataset

- Average topical representation for each rating:
  - Using features can find more regular patterns
    - Smoother changes of the trend from negative topics to positive ones
    - No ordering
Results on Large Dataset

Predictive R2:

![Graph showing Predictive R2 for different training data sizes varying from 500 to 2500 samples. The graph compares the performance of CTC, CTRF, and MedLDA methods. The y-axis represents Predictive R2 values ranging from 0.44 to 0.60. The x-axis represents Training data size in increments of 500 samples.]
Results on Large Dataset

Training & testing time:

- ~10 times faster
- ~100 times faster
Learning compact image representation

- Learn structured dictionary for encoding high-dimensional high-level image features

Joint with:
Li-jia Li, Hao Su, Eric Xing, & Li Fei-Fei
Experimental results

**Classification**

![Classification on UIUC Sports Event Dataset](image)

![Classification on MIT Indoor Dataset](image)

**Retrieval**

![Content Based Image Retrieval](image)

~40 times compression
**Summary so far**

- **Sparse topical coding** – an alternative topic model
  - a hierarchical sparse coding model performs regularized projection to convex cone with learned topical bases
Summary so far

- **Sparse topical coding** – an alternative topic model
  - Can be generalized to incorporate rich features
  - Can be applied to deal with images and videos

- the coordinate descent algorithm is available in C++
  [http://www.ml-thu.net/~jun/stc.html](http://www.ml-thu.net/~jun/stc.html)
More Applications

- Sparse representations for image restoration
- Sparse representations for video restoration
Image Denoising

\[ y \overset{\text{measurements}}{=} x_{\text{orig}} + w \overset{\text{noise}}{=} \]
Sparse representations for image restoration

General form:

\[
\begin{pmatrix}
x
\end{pmatrix}
= \begin{pmatrix}
d_1 & d_2 & \cdots & d_p
\end{pmatrix}
\begin{pmatrix}
\alpha[1] \\
\alpha[2] \\
\vdots \\
\alpha[p]
\end{pmatrix}
\]

\(x \in \mathbb{R}^m\)

Designed dictionaries

- Wavelets, Curvelets, Wedgelets, Bandlets, …lets

Learned dictionaries of patches [Olshausen & Field, 1997]

\[
\min_{\alpha_i, D \in \mathcal{C}} \sum_i \frac{1}{2} ||x_i - D\alpha_i||_2^2 + \lambda \psi(\alpha_i)
\]

\text{reconstruction} + \text{sparsity}
Sparse representations for image restoration

- Dictionary trained on a noisy version of the image boat [Elad & Aharon, 2006]
Sparse representations for image restoration

- Inpainting a 12-Mpixel photograph via online learning algorithm [Mairal et al., 2010]
Sparse representations for image restoration

Inpainting a 12-Mpixel photograph via online learning algorithm [Mairal et al., 2010]
Sparse representation for video restoration

Key ideas for video processing [Protter and Elad, 2009]

- Using a 3D dictionary
- Processing of many frames at the same time
- Dictionary propagation
Sparse representation for video restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]
Sparse representation for video restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]
Sparse representation for video restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]
Sparse representation for video restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]
Sparse representation for video restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]
Sparse representation for video restoration

Color video denoising, [Mairal, Sapiro, and Elad, 2008b]
Sparse representation for video restoration

- Color video denoising, [Mairal, Sapiro, and Elad, 2008b]
Sparse representation for video restoration

- Color video denoising, [Mairal, Sapiro, and Elad, 2008b]
Sparse representation for video restoration

- Color video denoising, [Mairal, Sapiro, and Elad, 2008b]
Sparse representation for video restoration

Color video denoising, [Mairal, Sapiro, and Elad, 2008b]
References


