Learning with Bregman Divergences

Machine Learning and Optimization

Inderjit S. Dhillon
University of Texas at Austin

Mathematical Programming in Data Mining and Machine Learning
Banff International Research Station, Canada
January 18, 2007

Joint work with Arindam Banerjee, Jason Davis, Joydeep Ghosh, Brian Kulis, Srujana Merugu and Suvrit Sra

Machine Learning: Problems

Unsupervised Learning

- Clustering: group a set of data objects
- Co-clustering: simultaneously partition data objects & features
- Matrix Approximation
  - SVD: low-rank approximation, minimizes Frobenius error
  - NNMA: low-rank non-negative approximation

Machine Learning: Problems

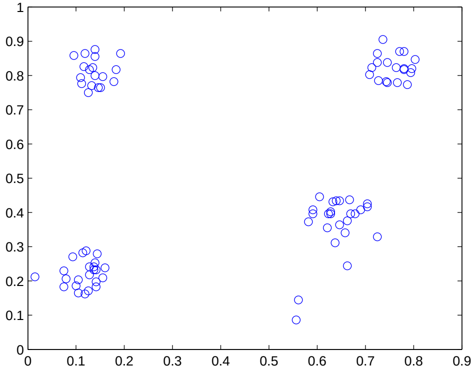
Unsupervised Learning

- Clustering: group a set of data objects
- Co-clustering: simultaneously partition data objects & features
- Matrix Approximation
  - SVD: low-rank approximation, minimizes Frobenius error
  - NNMA: low-rank non-negative approximation

Supervised Learning

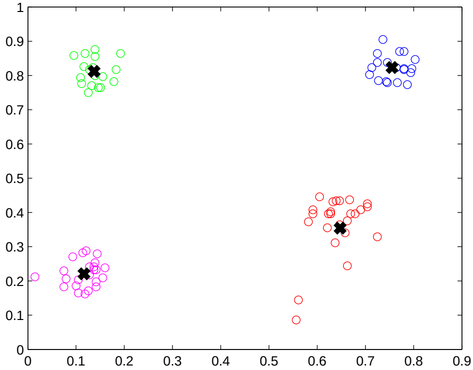
- Classification: k-nearest neighbor, SVMs, boosting, ...
  - Many classifiers rely on choice of distance measures
- Kernel Learning: used in "kernelized" algorithms
- Metric Learning: Information retrieval, Nearest neighbor searches

Example: Clustering



Goal: partition points into k clusters

Example: K-Means Clustering



Minimizes squared Euclidean distance from points to their cluster centroids

Example: K-Means Clustering

- Assumes a Gaussian noise model
  - Corresponds to squared Euclidean distance
- What if a different noise model is assumed?
  - Poisson, multinomial, exponential, etc.
- We will see: for every exponential family probability distribution, there exists a corresponding generalized distance measure

Distribution	**Distance Measure**
Spherical Gaussian	Squared Euclidean Distance
Multinomial	Kullback-Leibler Distance
Exponential	Itakura-Saito Distance

- Leads to generalizations of the k-means objective
  - **Bregman divergences** are the generalized distance measures

Background

Bregman Divergences: Definition

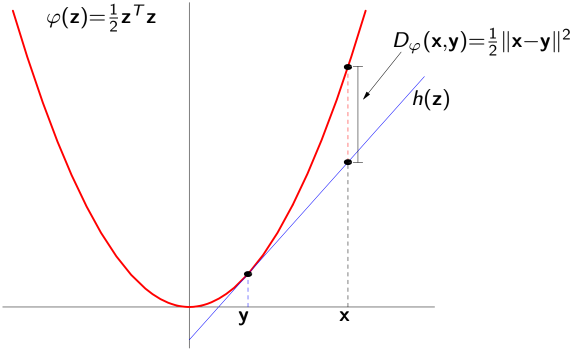
- Let $\varphi : S \to \mathbb{R}$ be a differentiable, strictly convex function of "Legendre type" ($S \subseteq \mathbb{R}^d$)
- The Bregman Divergence $D_\varphi : S \times \text{relint}(S) \to \mathbb{R}$ is defined as

$$D_\varphi(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - (\mathbf{x} - \mathbf{y})^T \nabla\varphi(\mathbf{y})$$

Bregman Divergences: Definition

- Let $\varphi : S \to \mathbb{R}$ be a differentiable, strictly convex function of "Legendre type" ($S \subseteq \mathbb{R}^d$)
- The Bregman Divergence $D_\varphi : S \times \text{relint}(S) \to \mathbb{R}$ is defined as

$$D_\varphi(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - (\mathbf{x} - \mathbf{y})^T \nabla\varphi(\mathbf{y})$$

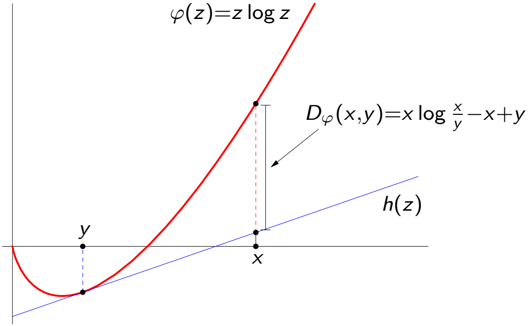


Squared Euclidean distance is a Bregman divergence

Bregman Divergences: Definition

- Let $\varphi : S \rightarrow \mathbb{R}$ be a differentiable, strictly convex function of "Legendre type" ($S \subseteq \mathbb{R}^d$)
- The Bregman Divergence $D_\varphi : S \times \text{relint}(S) \rightarrow \mathbb{R}$ is defined as

$$D_\varphi(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - (\mathbf{x} - \mathbf{y})^T \nabla\varphi(\mathbf{y})$$

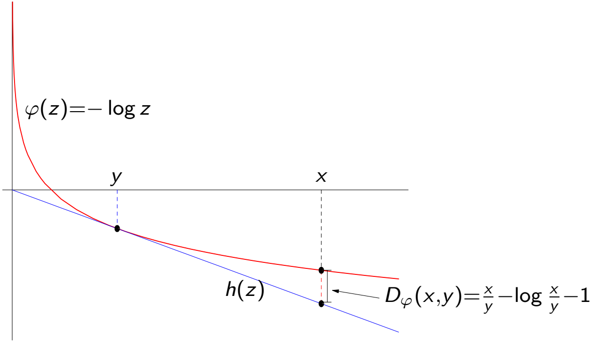


Relative Entropy (or KL-divergence) is another Bregman divergence

Bregman Divergences: Definition

- Let $\varphi : S \to \mathbb{R}$ be a differentiable, strictly convex function of "Legendre type" ($S \subseteq \mathbb{R}^d$)
- The Bregman Divergence $D_\varphi : S \times \text{relint}(S) \to \mathbb{R}$ is defined as

$$D_\varphi(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - (\mathbf{x} - \mathbf{y})^T \nabla\varphi(\mathbf{y})$$



Itakura-Saito Dist.(used in signal processing) is also a Bregman divergence

Bregman Divergences: Properties

- $D_\varphi(\mathbf{x}, \mathbf{y}) \geq 0$, and equals 0 iff $\mathbf{x} = \mathbf{y}$

Bregman Divergences: Properties

- $D_\varphi(\mathbf{x}, \mathbf{y}) \geq 0$, and equals 0 iff $\mathbf{x} = \mathbf{y}$
- Not a metric (symmetry, triangle inequality do not hold)

Bregman Divergences: Properties

- $D_\varphi(\mathbf{x}, \mathbf{y}) \geq 0$, and equals 0 iff $\mathbf{x} = \mathbf{y}$
- Not a metric (symmetry, triangle inequality do not hold)
- Strictly convex in 1^{st} argument, but (in general) not in 2^{nd}

Bregman Divergences: Properties

- $D_\varphi(\mathbf{x}, \mathbf{y}) \geq 0$, and equals 0 iff $\mathbf{x} = \mathbf{y}$
- Not a metric (symmetry, triangle inequality do not hold)
- Strictly convex in 1^{st} argument, but (in general) not in 2^{nd}
- Three-point property generalizes the "Law of cosines":

$$D_\varphi(\mathbf{x}, \mathbf{y}) = D_\varphi(\mathbf{x}, \mathbf{z}) + D_\varphi(\mathbf{z}, \mathbf{y}) - (\mathbf{x} - \mathbf{z})^T(\nabla\varphi(\mathbf{y}) - \nabla\varphi(\mathbf{z}))$$

Projections

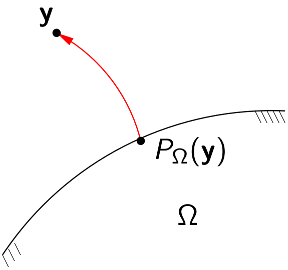
- "Bregman projection" of $\mathbf{y}$ onto a convex set Ω,

$$P_\Omega(\mathbf{y}) = \operatorname*{argmin}_{\boldsymbol{\omega}\in\Omega} D_\varphi(\boldsymbol{\omega}, \mathbf{y})$$

Projections

- "Bregman projection" of $\mathbf{y}$ onto a convex set Ω,

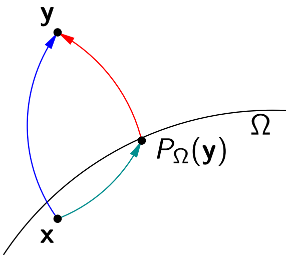
$$P_\Omega(\mathbf{y}) = \operatorname*{argmin}_{\boldsymbol{\omega} \in \Omega} D_\varphi(\boldsymbol{\omega}, \mathbf{y})$$



Projections

- "Bregman projection" of $\mathbf{y}$ onto a convex set Ω,

$$P_\Omega(\mathbf{y}) = \underset{\boldsymbol{\omega}\in\Omega}{\operatorname{argmin}}\, D_\varphi(\boldsymbol{\omega}, \mathbf{y})$$



- Generalized Pythagorean Theorem:

$$D_\varphi(\mathbf{x}, \mathbf{y}) \geq D_\varphi(\mathbf{x}, P_\Omega(\mathbf{y})) + D_\varphi(P_\Omega(\mathbf{y}), \mathbf{y})$$

When Ω is an affine set, the above holds with equality

Bregman's original work

- L. M. Bregman. "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming." *USSR Computational Mathematics and Physics*, 7:200-217, 1967.
  - Problem:

$$\min \varphi(\mathbf{x}) \qquad \text{subject to} \qquad \mathbf{a}_i^T \mathbf{x} = b_i, \;\; i = 0, \ldots, m-1$$

  - Bregman's cyclic projection method:
    - Start with appropriate $\mathbf{x}^{(0)}$. Compute $\mathbf{x}^{(t+1)}$ to be the Bregman projection of $\mathbf{x}^{(t)}$ onto the i-th hyperplane ($i = t \mod m$) for $t = 0, 1, 2, \ldots$
  - Converges to globally optimal solution. This cyclic projection method can be extended to halfspace and convex constraints, where each projection is followed by a correction.

Question: What role do Bregman divergences play in machine learning?

THE RELAXATION METHOD OF FINDING THE COMMON POINT OF CONVEX SETS AND ITS APPLICATION TO THE SOLUTION OF PROBLEMS IN CONVEX PROGRAMMING*

L.M. BREGMAN

Leningrad

(*Received* 20 *May* 1966)

IN this paper we consider an iterative method of finding the common point of convex sets. This method can be regarded as a generalization of the methods discussed in [1 - 4]. Apart from problems which can be reduced to finding some point of the intersection of convex sets, the method considered can be applied to the approximate solution of problems in linear and convex programming.

1. The problem of finding the common point of convex sets

Suppose we are given in a linear topological space X some family of closed convex sets A_i, $i \in I$, where I is some set of indices. We shall assume that $B = \bigcap_{i \in I} A_i$ is not empty. It is required to find some point of the intersection of the sets A_i.

Let $S \subset X$ be some convex set such that $S \cap B \neq \Lambda$.

Let us consider the function $D(x, y)$, defined over $S \times S$, and satisfying the following conditions.

I. $D(x, y) \geqslant 0$, $D(x, y) = 0$ if and only if $x = y$.

* *Zh. vȳchisl. Mat. mat. Fiz.* 7, 3, 620 - 631, 1967.

Exponential Families of Distributions

- **Definition**: A regular exponential family is a family of probability distributions on $\mathbb{R}^d$ with density function parameterized by $\boldsymbol{\theta}$:

$$p_\psi(\mathbf{x} \,|\, \boldsymbol{\theta}) = \exp\{\mathbf{x}^T\boldsymbol{\theta} - \psi(\boldsymbol{\theta}) - g_\psi(\mathbf{x})\}$$

  ψ is the so-called *cumulant function*, and is a convex function of Legendre type

Exponential Families of Distributions

- **Definition**: A regular exponential family is a family of probability distributions on $\mathbb{R}^d$ with density function parameterized by $\boldsymbol{\theta}$:

$$p_\psi(\mathbf{x} \mid \boldsymbol{\theta}) = \exp\{\mathbf{x}^T\boldsymbol{\theta} - \psi(\boldsymbol{\theta}) - g_\psi(\mathbf{x})\}$$

  ψ is the so-called *cumulant function*, and is a convex function of Legendre type

- **Example**: spherical Gaussians parameterized by mean $\boldsymbol{\mu}$ (& fixed variance σ):

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{x}-\boldsymbol{\mu}\|^2\right\} \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left\{\mathbf{x}^T\left(\frac{\boldsymbol{\mu}}{\sigma^2}\right) - \frac{\sigma^2}{2}\left(\frac{\boldsymbol{\mu}}{\sigma^2}\right)^2 - \frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2}\right\} \end{aligned}$$

Thus $\boldsymbol{\theta} = \dfrac{\boldsymbol{\mu}}{\sigma^2}$, and $\psi(\boldsymbol{\theta}) = \dfrac{\sigma^2}{2}\boldsymbol{\theta}^2$

Exponential Families of Distributions

- **Definition**: A regular exponential family is a family of probability distributions on $\mathbb{R}^d$ with density function parameterized by $\boldsymbol{\theta}$:

$$p_\psi(\mathbf{x} \,|\, \boldsymbol{\theta}) = \exp\{\mathbf{x}^T\boldsymbol{\theta} - \psi(\boldsymbol{\theta}) - g_\psi(\mathbf{x})\}$$

  ψ is the so-called *cumulant function*, and is a convex function of Legendre type

- **Example**: spherical Gaussians parameterized by mean $\boldsymbol{\mu}$ (& fixed variance σ):

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{x}-\boldsymbol{\mu}\|^2\right\} \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left\{\mathbf{x}^T\left(\frac{\boldsymbol{\mu}}{\sigma^2}\right) - \frac{\sigma^2}{2}\left(\frac{\boldsymbol{\mu}}{\sigma^2}\right)^2 - \frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2}\right\} \end{aligned}$$

$$\text{Thus } \boldsymbol{\theta} = \frac{\boldsymbol{\mu}}{\sigma^2}, \qquad \text{and} \qquad \psi(\boldsymbol{\theta}) = \frac{\sigma^2}{2}\boldsymbol{\theta}^2$$

- **Note:** Gaussian distribution $\longleftrightarrow$ Squared Loss

Example: Poisson Distribution

- Poisson Distribution:

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \mathbb{Z}_+$$

- The Poisson Distribution is a member of the exponential family
- Is there a Divergence associated with the Poisson Distribution?

Example: Poisson Distribution

- Poisson Distribution:

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \mathbb{Z}_+$$

- The Poisson Distribution is a member of the exponential family
- Is there a Divergence associated with the Poisson Distribution?

- YES — $p(x)$ can be written as

$$p(x) \quad = \quad \exp\{-D_\varphi(x, \mu) - g_\varphi(x)\},$$

where D_φ is the Relative Entropy, i.e., $D_\varphi(x, \mu) = x \log\left(\frac{x}{\mu}\right) - x + \mu$

Example: Poisson Distribution

- Poisson Distribution:

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \mathbb{Z}_+$$

- The Poisson Distribution is a member of the exponential family
- Is there a Divergence associated with the Poisson Distribution?

- YES — $p(x)$ can be written as

$$p(x) \quad = \quad \exp\{-D_\varphi(x, \mu) - g_\varphi(x)\},$$

where D_φ is the Relative Entropy, i.e., $D_\varphi(x, \mu) = x \log\left(\frac{x}{\mu}\right) - x + \mu$

- **Implication:** Poisson distribution $\longleftrightarrow$ Relative Entropy

Example: Exponential Distribution

- Exponential Distribution:

$$p(x) = \lambda \exp\{-\lambda x\}$$

- The Exponential Distribution is a member of the exponential family
- Is there a Divergence associated with the Exponential Distribution?

Example: Exponential Distribution

- Exponential Distribution:

$$p(x) = \lambda \exp\{-\lambda x\}$$

- The Exponential Distribution is a member of the exponential family
- Is there a Divergence associated with the Exponential Distribution?

- YES — $p(x)$ can be written as

$$p(x) \;=\; \exp\{-D_\varphi(x,\mu) - h(x)\},$$

where D_φ is the Itakura-Saito Distance, i.e., $D_\varphi(x,\mu) = \frac{x}{\mu} - \log\frac{x}{\mu} - 1$

Example: Exponential Distribution

- Exponential Distribution:

$$p(x) = \lambda \exp\{-\lambda x\}$$

- The Exponential Distribution is a member of the exponential family
- Is there a Divergence associated with the Exponential Distribution?

- YES — $p(x)$ can be written as

$$p(x) \quad = \quad \exp\{-D_{\varphi}(x, \mu) - h(x)\},$$

where D_{φ} is the Itakura-Saito Distance, i.e., $D_{\varphi}(x, \mu) = \frac{x}{\mu} - \log \frac{x}{\mu} - 1$

- **Implication:** Exponential distribution $\longleftrightarrow$ Itakura-Saito Dist.

Bregman Divergences and the Exponential Family

Theorem

Suppose that φ and ψ are conjugate Legendre functions. Let D_φ be the Bregman divergence associated with φ, and let $p_\psi(\cdot \,|\, \boldsymbol{\theta})$ be a member of the regular exponential family with cumulant function ψ. Then

$$p_\psi(\mathbf{x} \,|\, \boldsymbol{\theta}) = \exp\{-D_\varphi(\mathbf{x}, \boldsymbol{\mu}(\boldsymbol{\theta})) - g_\varphi(\mathbf{x})\},$$

where g_φ is a function uniquely determined by φ.

- Thus there is unique Bregman divergence associated with every member of the exponential family
- **Implication:** Member of Exponential Family $\longleftrightarrow$ unique Bregman Divergence.

Machine Learning Applications

Clustering with Bregman Divergences

- Let $\mathbf{a}_1, \ldots, \mathbf{a}_n$ be data vectors
- Goal: Divide data into k disjoint partitions $\gamma_1, \ldots, \gamma_k$
- Objective function for Bregman clustering:

$$\min_{\gamma_1,\ldots,\gamma_k} \sum_{h=1}^{k} \sum_{\mathbf{a}_i \in \gamma_h} D_\varphi(\mathbf{a}_i, \mathbf{y}_h),$$

  where $\mathbf{y}_h$ is the representative of the h-th partition
- **Lemma.** Arithmetic mean is the optimal representative for all D_φ:

$$\boldsymbol{\mu}_h \equiv \frac{1}{|\gamma_h|} \sum_{\mathbf{a}_i \in \gamma_h} \mathbf{a}_i \;=\; \operatorname*{argmin}_{\mathbf{x}} \sum_{\mathbf{a}_i \in \gamma_h} D_\varphi(\mathbf{a}_i, \mathbf{x})$$

  - Reverse implication also holds
- Algorithm: `KMeans`-type iterative re-partitioning algorithm monotonically decreases objective

Co-Clustering with Bregman Divergences

- Let $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ be an $m \times n$ data matrix
- Goal: partition A into k row clusters and ℓ column clusters
- How do we judge the quality of co-clustering?

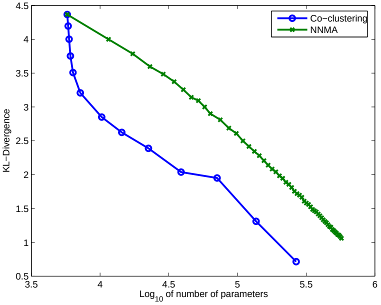
Co-Clustering with Bregman Divergences

- Let $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ be an $m \times n$ data matrix
- Goal: partition A into k row clusters and ℓ column clusters
- How do we judge the quality of co-clustering?
- Use quality of "associated" matrix approximation
  - Associate matrix approximation using the Minimum Bregman Information (MBI) principle
- Objective: Find optimal co-clustering $\leftrightarrow$ optimal MBI approximation

Co-Clustering with Bregman Divergences

- Let $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ be an $m \times n$ data matrix
- Goal: partition A into k row clusters and ℓ column clusters
- How do we judge the quality of co-clustering?
- Use quality of "associated" matrix approximation
  - Associate matrix approximation using the Minimum Bregman Information (MBI) principle
- Objective: Find optimal co-clustering $\leftrightarrow$ optimal MBI approximation
- Example: Information-Theoretic Co-Clustering
  - Measures approximation error using relative entropy

Co-Clustering as Matrix Approximation

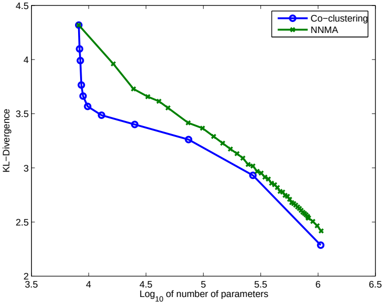


Error of approximation vs. number of parameters

$$M = 5471, N = 300$$

NNMA approximation computed using Lee & Seung's algorithm

Co-Clustering as Matrix Approximation



Error of approximation vs. number of parameters

$$M = 4303, N = 3891$$

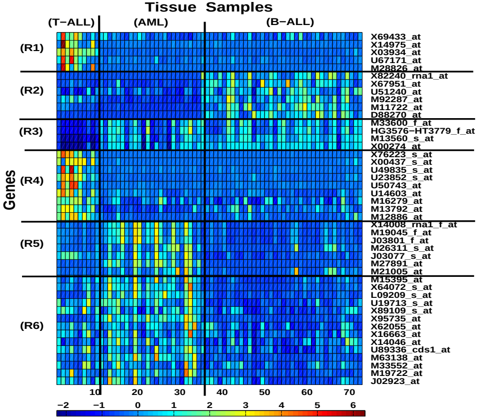
NNMA approximation computed using Lee & Seung's algorithm

Co-Clustering Applied to Bioinformatics

- Gene Expression Leukemia data
- Matrix contains expression levels of genes in different tissue samples

Co-Clustering Applied to Bioinformatics

- Gene Expression Leukemia data
- Matrix contains expression levels of genes in different tissue samples
- Co-clustering recovers cancer samples & functionally related genes



Learning Over Matrix Inputs

- Many problems in machine learning require optimization over symmetric matrices
- Kernel learning: find a kernel matrix that satisfies a set of constraints
  - Support vector machines
  - Semi-supervised graph clustering via kernels
- Distance metric learning: find a Mahalanobis distance metric
  - Information retrieval
  - k-Nearest neighbor classification

Learning Over Matrix Inputs

- Many problems in machine learning require optimization over symmetric matrices
- Kernel learning: find a kernel matrix that satisfies a set of constraints
  - Support vector machines
  - Semi-supervised graph clustering via kernels
- Distance metric learning: find a Mahalanobis distance metric
  - Information retrieval
  - k-Nearest neighbor classification
- Bregman divergences can be naturally extended to matrix-valued inputs

Bregman Matrix Divergences

- Let
  - $\mathcal{H}$: space of $N \times N$ Hermitian matrices
  - $\boldsymbol{\lambda} : \mathcal{H} \to \mathbb{R}^N$ be the eigenvalue map
  - $\varphi : \mathbb{R}^N \to \mathbb{R}$ be a convex function of Legendre type
  - $\hat{\varphi} = \varphi \circ \boldsymbol{\lambda}$
- Define

$$D_{\hat{\varphi}}(A, B) = \hat{\varphi}(X) - \hat{\varphi}(Y) - \operatorname{trace}((\nabla \hat{\varphi}(Y))^*(X - Y))$$

- Squared Frobenius norm: $\hat{\varphi}(X) = \|X\|_F^2$. Then

$$D_{\hat{\varphi}}(X, Y) = \frac{1}{2}\|X - Y\|_F^2$$

  - Used in many nearness problems

Bregman Matrix Divergences

- von Neumann Divergence: For $X \succeq 0$, $\hat{\varphi}(X) = \text{trace}(X \log X)$. Then

$$D_{\hat{\varphi}}(X, Y) = \text{trace}(X \log X - X \log Y - X + Y)$$

  - also called quantum relative entropy

Bregman Matrix Divergences

- von Neumann Divergence: For $X \succeq 0$, $\hat{\varphi}(X) = \text{trace}(X \log X)$. Then

$$D_{\hat{\varphi}}(X, Y) = \text{trace}(X \log X - X \log Y - X + Y)$$

  - also called quantum relative entropy

- LogDet divergence: For $X \succ 0$, $\hat{\varphi}(X) = -\log\det X$. Then

$$D_{\hat{\varphi}}(X, Y) = \text{trace}(XY^{-1}) - \log\det(XY^{-1}) - N$$

  - **Interesting Connection:** The differential relative entropy between two equal-mean Gaussians with covariance matrices X and Y EXACTLY equals the LogDet divergence between X and Y

Low-Rank Kernel Learning

- Learn a low-rank spd matrix that satisfies given constraints:

$$
\begin{aligned}
\min_{K} \quad & D_{\hat{\varphi}}(K, K_0) \\
\text{subject to} \quad & \text{trace}(KA_i) \leq b_i, \; 1 \leq i \leq c \\
& \text{rank}(K) \leq r \\
& K \succeq 0
\end{aligned}
$$

Low-Rank Kernel Learning

- Learn a low-rank spd matrix that satisfies given constraints:

$$\begin{aligned} \min_{K} \quad & D_{\hat{\varphi}}(K, K_0) \\ \text{subject to} \quad & \text{trace}(KA_i) \leq b_i, \; 1 \leq i \leq c \\ & \text{rank}(K) \leq r \\ & K \succeq 0 \end{aligned}$$

- Problem is non-convex due to rank constraint

Preserving Range Space

Lemma

Suppose φ is separable, i.e., $\varphi(\mathbf{x}) = \sum_i \varphi_s(x_i)$. Let the spectral decompositions of X and Y be $X = V\Lambda V^T$ and $Y = U\Theta U^T$. Then

$$D_{\hat{\varphi}}(X, Y) = \sum_i \sum_j (\mathbf{v}_i^T \mathbf{u}_j)^2 D_{\varphi_s}(\lambda_i, \theta_j).$$

- Example: LogDet Divergence can be written as

$$D_{LogDet}(X, Y) = \sum_i \sum_j (\mathbf{v}_i^T \mathbf{u}_j)^2 \left(\frac{\lambda_i}{\theta_j} - \log \frac{\lambda_i}{\theta_j} - 1 \right)$$

- **Corollary 1:** $D_{vN}(X, Y)$ finite iff $\text{range}(X) \subseteq \text{range}(Y)$
- **Corollary 2:** $D_{LogDet}(X, Y)$ finite iff $\text{range}(X) = \text{range}(Y)$

Low-Rank Kernel Learning

- **Implication:** $\text{rank}(K) \leq \text{rank}(K_0)$ for vN-divergence and $\text{rank}(K) = \text{rank}(K_0)$ for LogDet divergence
- Adapt Bregman's algorithm to solve the problem

$$\begin{array}{ll} \min_{K} & D_{\hat{\varphi}}(K, K_0) \\ \text{subject to} & \text{trace}(KA_i) \leq b_i,\ 1 \leq i \leq c \end{array}$$

- Algorithm works on *factored* forms of the kernel matrix
- Bregman projections onto a rank-one constraint can be computed in $O(r^2)$ time for both divergences

Details

- LogDet divergence
  - Projection can be easily computed in closed-form
  - Iterate is updated using Sherman-Morrison formula
  - Requires $O(r^2)$ Cholesky decomposition of $I + \alpha \mathbf{x}\mathbf{x}^T$
- von Neumann divergence
  - Projection computed by custom non-linear solver with quadratic convergence
  - Iterate is updated using eigenvalue decomposition of $I + \alpha \mathbf{x}\mathbf{x}^T$
  - Requires $O(r^2)$ update using fast multipole method
- Largest problem size handled: $n = 20,000$ with $r = 16$
- Useful for learning low-rank kernels for support vector machines, semi-supervised clustering, etc.

Information-Theoretic Metric Learning

- Problem: Learn a Mahalanobis metric

$$d_X(\mathbf{y}_1, \mathbf{y}_2) = (\mathbf{y}_1 - \mathbf{y}_2)^T X (\mathbf{y}_1 - \mathbf{y}_2)$$

that satisfies given pairwise distance constraints

- The following problems are equivalent:

Metric Learning		Kernel Learning
$$\begin{aligned} \min_X \quad & KL(p(\mathbf{y};\boldsymbol{\mu},X)\|p(\mathbf{y};\boldsymbol{\mu},I)) \\ \text{s.t.} \quad & d_X(\mathbf{y}_i,\mathbf{y}_j) \le U, (i,j) \in S \\ & d_X(\mathbf{y}_i,\mathbf{y}_j) \ge L, (i,j) \in D \\ & X \succeq 0 \end{aligned}$$	$\equiv$	$$\begin{aligned} \min_K \quad & D_{\hat{\varphi}}(K, K_0) \\ \text{s.t.} \quad & \text{trace}(KA_i) \le b_i \\ & \text{rank}(K) \le r \\ & K \succeq 0 \end{aligned}$$

- where the connection is that $K_0 = Y^T Y$, $K = Y^T X Y$ and $r = m$
- Note that K_0 and K are low-rank when $n > m$

Challenges

Algorithms

- Bregman's method is simple, but suffers from slow convergence
- Interior point methods?
- Numerical stability?

Challenges

Algorithms

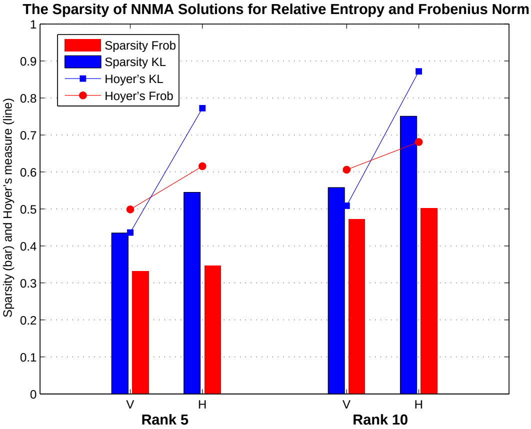
- Bregman's method is simple, but suffers from slow convergence
- Interior point methods?
- Numerical stability?

Choosing an appropriate Bregman Divergence

- Noise models are not always available
- How to choose the best Bregman divergence?

What Bregman Divergence to use?

- NNMA approximation: $\mathbf{A} \approx \mathbf{VH}$
- Some divergences might preserve sparsity better than others



References

Clustering

"Clustering with Bregman Divergences", A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Journal of Machine Learning Research, vol. 6, pages 1705-1749, October 2005.

"A Generalized Maximum Entropy Approach to Bregman Co-Clustering and Matrix Approximations", A. Banerjee, I. S. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. ACM Conference on Knowledge Discovery and Data Mining(KDD), pages 509-514, August 2004.

"Co-clustering of Human Cancer Microarrays using Minimum Squared Residue Co-clustering", H. Cho and I. S. Dhillon. submitted for publication, 2006.

"Differential Entropic Clustering of Multivariate Gaussians", J. V. Davis and I. S. Dhillon. Neural Information Processing Systems Conference (NIPS), December 2006.

NNMA

"Generalized Nonnegative Matrix Approximations with Bregman Divergences", I. S. Dhillon and S. Sra. Neural Information Processing Systems Conference (NIPS), pages 283-290, Vancouver Canada, December 2005.

"Fast Newton-type Methods for the Least Squares Nonnegative Matrix Approximation Problem", D. Kim, S. Sra, and I. S. Dhillon. SIAM International Conference on Data Mining, April 2007.

References

Kernel & Metric Learning

"Learning Low-Rank Kernel Matrices", B. Kulis, M. A. Sustik, and I. S. Dhillon. International Conference on Machine Learning (ICML), pages 505-512, July 2006.

"Matrix Exponentiated Gradient Updates for Online Learning and Bregman Projection", K. Tsuda, G. Ratsch, and M. Warmuth. Journal of Machine Learning Research, vol. 6, pages 995-1018, December 2004.

"Information-Theoretic Metric Learning", J. V. Davis, B. Kulis, S. Sra, and I. S. Dhillon. Neural Information Processing Systems Workshop on *Learning to Compare Examples*, 2006.

Classification

"Logistic Regression, AdaBoost and Bregman Distances", M. Collins, R. Schapire, and Y. Singer. Machine Learning, vol. 48, pages 253-285, 2000.

Expository Article

"Matrix Nearness Problems using Bregman Divergences", I. S. Dhillon and J. A. Tropp, To appear in SIAM Journal on Matrix Analysis and its Applications, 2007.