Clustering with Bregman Divergences
Banerjee, Merugu, Dhillon and Ghosh, JMLR 2005

Nisheeth Srivastava

Dept of Computer Science
University of Minnesota

September 14, 2007
Introduction

Standard EM algorithm
Bregman Divergences
Regular exponential families

An equivalence relationship

A Legendre dual
Exponential families to Bregman divergences
Bregman divergences to regular exponential families

Clustering and Mixture Modeling

Motivation
The Bregman advantage
Bregman k-means

Bregman Information

Information-theoretic clustering

Nuisance parameters

Experiments
Open questions
Outline

Introduction

Standard EM algorithm
  Bregman Divergences
  Regular exponential families

An equivalence relationship
  A Legendre dual
  Exponential families to Bregman divergences
  Bregman divergences to regular exponential families

Clustering and Mixture Modeling

Motivation
  The Bregman advantage
  Bregman k-means

Bregman Information
  Information-theoretic clustering

Nuisance parameters
Iterative Expectation-Maximization

- Initialize \( \{\theta_h, \pi_h\}_{h=1}^k \)
- The expectation step
  
  \[
  p(h|\mathbf{x}_i) \leftarrow \frac{\pi_h p(\psi, \theta_h)(\mathbf{x}_i)}{\sum_{h'=1}^k \pi_{h'} p(\psi, \theta_{h'})(\mathbf{x}_i)}
  \]

- The Maximization step
  
  \[
  \pi_h \leftarrow \frac{1}{n} p(h | \mathbf{x}_i)
  \]
  
  \[
  \theta_h \leftarrow \arg\max_{\theta} \sum_{i=1}^n \log(p(\psi, \theta_h)(\mathbf{x}_i)) p(h | \mathbf{x}_i)
  \]

- until convergence
Problems with EM

- Max likelihood estimation
  - Local minima
  - M step computationally intensive
Problems with EM

- Max likelihood estimation
  - Local minima
  - M step computationally intensive

- Find a good replacement for
  \[ \theta_h \leftarrow \text{argmax}_\theta \sum_{i=1}^n \log(p(\psi,\theta_h)(x_i))p(h \mid x_i) \]
Problems with EM

- Max likelihood estimation
  - Local minima
  - M step computationally intensive
- Find a good replacement for
  \[ \theta_h \leftarrow \arg \max_{\theta} \sum_{i=1}^{n} \log(p(\psi, \theta_h)(x_i)) p(h \mid x_i) \]
Outline

Introduction
- Standard EM algorithm
- Bregman Divergences
- Regular exponential families

An equivalence relationship
- A Legendre dual
- Exponential families to Bregman divergences
- Bregman divergences to regular exponential families

Clustering and Mixture Modeling
- Motivation
- The Bregman advantage
- Bregman k-means

Bregman Information
- Information-theoretic clustering

Nuisance parameters
Bregman Divergence

\[ d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla \phi(y) \rangle \]
Bregman Divergence

\[ d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla \phi(y) \rangle \]

- Some properties
Bregman Divergence

- \( d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla \phi(y) \rangle \)
- Some properties
  - \( \phi \) is strictly convex and is defined on convex set \( S \subseteq \mathbb{R}^d \)
Bregman Divergence

- $d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla \phi(y) \rangle$

Some properties

- $\phi$ is strictly convex and is defined on convex set $S \subseteq \mathbb{R}^d$
- $d_\phi(x, y) \geq 0$, with equality only when $x = y$
Bregman Divergence

- \( d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla \phi(y) \rangle \)
- Some properties
  - \( \phi \) is strictly convex and is defined on convex set \( S \subseteq \mathbb{R}^d \)
  - \( d_\phi(x, y) \geq 0 \), with equality only when \( x = y \)
  - If \( \phi(x) = \phi_0(x) + \langle b, x \rangle + c \), \( d_\phi(x, y) = d_{\phi_0}(x, y) \)
An example

\[ d_\phi(x, y) = \sum_{j=1}^{d} p_j \log_2 p_j - \sum_{j=1}^{d} q_j \log_2 q_j - \langle p - q, \nabla \phi(q) \rangle \]

\[ = \sum_{j=1}^{d} p_j \log_2 p_j - \sum_{j=1}^{d} q_j \log_2 q_j - \]

\[ \sum_{j=1}^{d} (p_j - q_j)(\log_2 q_j + \log_2 e) \]

\[ = \sum_{j=1}^{d} p_j \log_2 \left( \frac{p_j}{q_j} \right) - \log_2 e \sum_{j=1}^{d} (p_j - q_j) \]

\[ = KL(p \parallel q) \]
Outline

Introduction
Standard EM algorithm
Bregman Divergences
Regular exponential families
An equivalence relationship
A Legendre dual
Exponential families to Bregman divergences
Bregman divergences to regular exponential families
Clustering and Mixture Modeling
Motivation
The Bregman advantage
Bregman k-means
Bregman Information
Information-theoretic clustering
Nuisance parameters

Nisheeth Srivastava
Clustering with Bregman Divergences
Definition

$p(\psi, \theta)(x) = \exp(\langle x, \theta \rangle - \psi(\theta)) p_0(x) \quad \forall x \in \mathbb{R}^d$
Definition

\[ p(\psi, \theta)(x) = \exp(\langle x, \theta \rangle - \psi(\theta)) p_0(x) \quad \forall x \in \mathbb{R}^d \]

where,

- \( x \) is a minimal natural statistic for the family
- Parameter space \( \Theta \) is open
Definition

\[ p(\psi, \theta)(x) = \exp(\langle x, \theta \rangle - \psi(\theta)) p_0(x) \quad \forall x \in \mathbb{R}^d \]

where,

- \( x \) is a minimal natural statistic for the family
- Parameter space \( \Theta \) is open

Also,

- Cumulant function \( \psi \) unique for a family
- \( (\Theta, \psi) \) is a Legendre function
The expectation parameter

Definition
Given a regular exponential family density \( p_{(\psi, \theta)} \) specified by the natural parameter \( \theta \in \Theta \), the expectation of \( X \) with respect to \( p \) is called the *expectation parameter*, and is given by

\[
\mu = \mu(\theta) = \int_{\mathbb{R}^d} x p_{(\psi, \theta)}(x) \, dx
\]
Outline

Introduction
  Standard EM algorithm
  Bregman Divergences
  Regular exponential families

An equivalence relationship
  A Legendre dual
  Exponential families to Bregman divergences
  Bregman divergences to regular exponential families

Clustering and Mixture Modeling
  Motivation
  The Bregman advantage
  Bregman k-means

Bregman Information
  Information-theoretic clustering

Nuisance parameters
A legendary duel

Figure: The duel of Faerûn: Forgotten Realms
Legendre duality

Definition
If $\psi$ be a real-valued function on $\mathbb{R}^d$, its conjugate function $\psi^*$ is given by

$$\psi^*(t) = \sup_{\theta \in \text{dom}(\psi)} \{ \langle t, \theta \rangle - \psi(\theta) \}$$

Theorem
If $(\Theta, \psi)$ is a convex function of the Legendre type then

1. The gradient function $\nabla \psi : \Theta \mapsto \Theta^*$ is a one-to-one function from the open convex set $\Theta$ to the open convex set $\Theta^*$
2. The gradient functions $\nabla \psi$ and $\nabla \psi^*$ are continuous and $\nabla \psi^* = (\nabla \psi)^{-1}$
Expectation/natural parameter duality

\[ \int p(\psi, \theta)(x) \, dx = 1 \]

Differentiating with respect to \( \theta \)

\[ \int \frac{\partial}{\partial \theta} \exp(\langle x, \theta \rangle - \psi(\theta)) \, p_0(x) \, dx = 0 \]

Then

\[ \mu = \mu(\theta) = \nabla \psi(\theta) \] (1)

Define conjugate of \( \psi \)

\[ \phi(\mu) = \sup_{\theta \in \Theta} \{ \langle \mu, \theta \rangle - \psi(\theta) \} \]
Dual space mapping

\[ \Theta \text{ and } \text{int(dom}(\phi)) \text{ will have the following mapping} \]

\[ \mu(\theta) = \nabla \psi(\theta) \quad \text{and} \quad \theta(\mu) = \nabla \phi(\mu) \quad (2) \]

Conjugate function can be expressed as

\[ \phi(\mu) = \langle \theta(\mu), \mu \rangle - \psi(\theta(\mu)), \quad \forall \mu \in \text{int(dom}(\phi)) \quad (3) \]
Outline

Introduction
  Standard EM algorithm
  Bregman Divergences
  Regular exponential families

An equivalence relationship
  A Legendre dual
  Exponential families to Bregman divergences
  Bregman divergences to regular exponential families

Clustering and Mixture Modeling
  Motivation
  The Bregman advantage
  Bregman k-means

Bregman Information
  Information-theoretic clustering

Nuisance parameters
A simple transformation

\[ \langle x, \theta \rangle - \psi(\theta) = (\langle \mu, \theta \rangle - \psi(\theta)) + \langle x - \mu, \theta \rangle \]
\[ = \phi(\mu) + \langle x - \mu, \nabla \phi(\mu) \rangle \]
\[ = -d_\phi(x, \mu) + \phi(x) \]

where,
\[ x \in \text{dom}(\phi), \ \theta \in \Theta \text{ and } \mu \in \text{int}(\text{dom}(\phi)) \]
A tricky alignment

$$\log(p_{(\psi,\theta)}(x)) = -d_\phi(x, \mu) + \log(b_\phi(x))$$

where,

$$b_\phi(x) = \exp(\phi(x)) p_0(x)$$
A tricky alignment

\[ \log(p_{\psi,\theta}(x)) = -d_\phi(x, \mu) + \log(b_\phi(x)) \]

where,

\[ b_\phi(x) = \exp(\phi(x)) p_0(x) \]

▶ We don’t know if \( l_\psi \) is identical with \( \text{dom}(\phi) \)
Theorem

Let $I_\psi$ be the set of instances that can be drawn following $p_{(\psi,\theta)}(x)$. Then $I_\psi \subseteq \text{dom}(\phi)$, where $\phi$ is the conjugate function of $\psi$. 
The main theorem

**Theorem**

Let \( p(\psi, \theta) \) be the probability density function of a regular exponential family distribution.

Then \( p(\psi, \theta) \) can be uniquely expressed as

\[
p(\psi, \theta)(x) = \exp(-d_\phi(x, \mu)) b_\phi(x), \quad \forall x \in \text{dom}(\phi)
\]
The main theorem

**Theorem**

Let \( p(\psi, \theta) \) be the probability density function of a regular exponential family distribution. Let \( \phi \) be the conjugate function of \( \psi \), so that \((\text{int(dom}(\phi)), \phi)\) is the Legendre dual of \((\Theta, \psi)\).

Then \( p(\psi, \theta) \) can be uniquely expressed as

\[
p(\psi, \theta)(x) = \exp(-d_\phi(x, \mu)) b_\phi(x), \quad \forall x \in \text{dom}(\phi)
\]
The main theorem

**Theorem**

Let $p(\psi,\theta)$ be the probability density function of a regular exponential family distribution. Let $\phi$ be the conjugate function of $\psi$, so that $(\text{int}(\text{dom}(\phi)), \phi)$ is the Legendre dual of $(\Theta, \psi)$. Let $\theta \in \Theta$ be the natural parameter and $\mu \in \text{int}(\text{dom}(\phi))$ be the corresponding expectation parameter.

Then $p(\psi,\theta)$ can be uniquely expressed as

$$p(\psi,\theta)(x) = \exp(-d_\phi(x, \mu)) b_\phi(x), \quad \forall x \in \text{dom}(\phi)$$
The main theorem

Theorem
Let $p_{(\psi, \theta)}$ be the probability density function of a regular exponential family distribution. Let $\phi$ be the conjugate function of $\psi$, so that $(\text{int}(\text{dom}(\phi)), \phi)$ is the Legendre dual of $(\Theta, \psi)$. Let $\theta \in \Theta$ be the natural parameter and $\mu \in \text{int}(\text{dom}(\phi))$ be the corresponding expectation parameter. Let $d_{\phi}$ be the Bregman divergence derived from $\phi$.

Then $p_{(\psi, \theta)}$ can be uniquely expressed as

$$p_{(\psi, \theta)}(x) = \exp(-d_{\phi}(x, \mu)) b_{\phi}(x), \quad \forall x \in \text{dom}(\phi)$$
Outline

Introduction
- Standard EM algorithm
- Bregman Divergences
- Regular exponential families

An equivalence relationship
- A Legendre dual
- Exponential families to Bregman divergences
- Bregman divergences to regular exponential families

Clustering and Mixture Modeling
- Motivation
- The Bregman advantage
- Bregman k-means

Bregman Information
- Information-theoretic clustering

Nuisance parameters
Validity of the converse

- Regular exponential families $\leftrightarrow$ Bregman divergence
Validity of the converse

- Regular exponential families $\leftrightarrow$ Bregman divergence
- Bregman divergence $\leftrightarrow$ regular exponential family ??
A friend in need

Theorem (Devinatz, 1955)

Let \( \Theta \subseteq \mathcal{R}^d \) be an open convex set. A necessary and sufficient condition that there exists a unique, bounded, non-negative measure \( \nu \) such that \( f : \Theta \mapsto \mathcal{R}_{++} \) can be represented as

\[
    f(\theta) = \int_{x \in \mathcal{R}^d} \exp(\langle x, \theta \rangle) \, d\nu(x)
\]  

is that \( f \) is continuous and exponentially convex.
Definition
Let $\psi$ be a strictly convex function and $\phi$ be its conjugate. Then the Bregman divergence $d_\phi$ derived from $\phi$ is a regular Bregman divergence.
Validity of the converse

- Regular exponential families $\mapsto$ Bregman divergence
Validity of the converse

- Regular exponential families $\leftrightarrow$ Bregman divergence
- Regular Bregman divergence $\leftrightarrow$ Regular exponential family
Outline

**Introduction**
- Standard EM algorithm
- Bregman Divergences
- Regular exponential families

**An equivalence relationship**
- A Legendre dual
- Exponential families to Bregman divergences
- Bregman divergences to regular exponential families

**Clustering and Mixture Modeling**
- Motivation
  - The Bregman advantage
  - Bregman k-means

**Bregman Information**
- Information-theoretic clustering
- Nuisance parameters

Nisheeth Srivastava

Clustering with Bregman Divergences
Two sides of a coin

- Soft clustering ↔ Finite mixture modeling
Two sides of a coin

- Soft clustering $\leftrightarrow$ Finite mixture modeling
- Clusters $\leftrightarrow$ Mixture components
- Cluster membership probability $\leftrightarrow$ Probability generated by mixture component
Bregman soft clustering

Definition
The Bregman soft clustering problem is defined as that of learning the maximum likelihood parameters $\Gamma = \{\theta_h, \pi_h\} \equiv \{\mu_h, \pi_h\}$ of a mixture model of the form,

$$p(x | \Gamma) =$$
Definition
The Bregman soft clustering problem is defined as that of learning the maximum likelihood parameters $\Gamma = \{\theta_h, \pi_h\} \equiv \{\mu_h, \pi_h\}$ of a mixture model of the form,

$$p(x \mid \Gamma) = \sum_{h=1}^{k} \pi_h p(\psi, \theta_h)(x)$$
Definition
The Bregman soft clustering problem is defined as that of learning the maximum likelihood parameters $\Gamma = \{\theta_h, \pi_h\} \equiv \{\mu_h, \pi_h\}$ of a mixture model of the form,

$$p(x | \Gamma) = \sum_{h=1}^{k} \pi_h p_{(\psi,\theta_h)}(x) = \sum_{h=1}^{k} \pi_h \exp(-d_\phi(x, \mu_h))b_\phi(x)$$
Outline

Introduction
- Standard EM algorithm
- Bregman Divergences
- Regular exponential families

An equivalence relationship
- A Legendre dual
- Exponential families to Bregman divergences
- Bregman divergences to regular exponential families

Clustering and Mixture Modeling
- Motivation
  - The Bregman advantage
  - Bregman k-means

Bregman Information
- Information-theoretic clustering
- Nuisance parameters
Iterative Expectation-Maximization

- Initialize $\{\theta_h, \pi_h\}_{h=1}^k$
- The expectation step
  
  for $i = 1$ to $n$
  
  for $h = 1$ to $k$
  
  $$p(h|x_i) \leftarrow \frac{\pi_h p(\psi, \theta_h)(x_i)}{\sum_{h'=1}^k \pi_{h'} p(\psi, \theta_{h'})(x_i)}$$
  
  end for
  
  end for

- The Maximization step
  
  for $h = 1$ to $k$
  
  $$\pi_h \leftarrow \frac{1}{n} p(h | x_i)$$
  
  $$\theta_h \leftarrow \arg\max_{\theta} \sum_{i=1}^n \log(p(\psi, \theta_h)(x_i)) p(h | x_i)$$
  
  end for

- until convergence
Natural to unnatural

Maximization step

\[ \theta_h = \arg\max_{\theta} \sum_{i=1}^{n} \log(p(\psi, \theta)(x_i)) p(h \mid x_i) \]

is equivalent to

\[ \mu_h = \arg\max_{\mu} \sum_{i=1}^{n} \log(b_\phi(x_i) \exp(-d_\phi(x_i, \mu))) p(h \mid x_i) \]

\[ = \arg\max_{\mu} \sum_{i=1}^{n} (\log(b_\phi(x_i)) - d_\phi(x_i, \mu)) p(h \mid x_i) \]

\[ = \arg\min_{\mu} \sum_{i=1}^{n} d_\phi(x_i, \mu) \frac{p(h \mid x_i)}{\sum_{i' = 1}^{n} p(h \mid x_{i'})} \]
A surprising result

Proposition

Let $X$ be a random variable in $\mathcal{X} = \{x_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$ following a positive probability measure $\nu$ such that $E_\nu[X] \in ri(S)$. Given a Bregman divergence $d_\phi : S \times ri(S) \rightarrow [0, \infty)$, the problem

$$\min_{s \in ri(S)} E_\nu[d_\phi(X, s)]$$

has a unique minimizer given by $s^\dagger = \mu = E_\nu[X]$. 

Nisheeth Srivastava
Clustering with Bregman Divergences
Bregman soft clustering

- Initialize $\{\theta_h, \pi_h\}_{h=1}^k$
- The expectation step
  
  $\text{for } i = 1 \text{ to } n \text{ do}$
  
  $\text{for } h = 1 \text{ to } k \text{ do}$
  
  $p(h|x_i) \leftarrow \frac{\pi_h p(\psi, \theta_h)(x_i)}{\sum_{h' = 1}^k \pi_{h'} p(\psi, \theta_{h'})(x_i)}$

  $\text{end for}$

  $\text{end for}$

- $\text{until convergence}$
Bregman soft clustering

- Initialize $\{\theta_h, \pi_h\}_{h=1}^k$
- The expectation step
  
  \[
  \text{for } i = 1 \text{ to } n \text{ do}
  \]
  \[
  \text{for } h = 1 \text{ to } k \text{ do}
  \]
  \[
  p(h|\mathbf{x}_i) \leftarrow \frac{\pi_h p_\psi(\theta_h)(\mathbf{x}_i)}{\sum_{h'=1}^k \pi_{h'} p_\psi(\theta_{h'})(\mathbf{x}_i)}
  \]
  \[
  \text{end for}
  \]
  \[
  \text{end for}
  \]
- The Maximization step
  
  \[
  \text{for } h = 1 \text{ to } k \text{ do}
  \]
  \[
  \pi_h \leftarrow \frac{1}{n} p(h|\mathbf{x}_i)
  \]
  \[
  \mu_h \leftarrow \frac{\sum_{i=1}^n p(h|\mathbf{x}_i) \mathbf{x}_i}{\sum_{i=1}^n p(h|\mathbf{x}_i)}
  \]
  \[
  \text{end for}
  \]
- until convergence
Outline

Introduction
Standard EM algorithm
Bregman Divergences
Regular exponential families
An equivalence relationship
A Legendre dual
Exponential families to Bregman divergences
Bregman divergences to regular exponential families

Clustering and Mixture Modeling
Motivation
The Bregman advantage
Bregman k-means

Bregman Information
Information-theoretic clustering
Nuisance parameters
Consider the update equation for the E-step

\[ p(h \mid x) = \frac{\pi_h \exp(-d_\phi(x, \mu_h))}{\sum_{h'=1}^{k} \pi_{h'} \exp(-d_\phi(x, \mu_{h'}))} \]
Consider the update equation for the E-step

\[ p(h | x) = \frac{\pi_h \exp(-d_\phi(x, \mu_h))}{\sum_{h' = 1}^{k} \pi_{h'} \exp(-d_\phi(x, \mu_{h'}))} \]

\[ d_{\beta \phi} = \beta d_\phi \]
Consider the update equation for the E-step

\[ p(h \mid x) = \frac{\pi_h \exp(-d_\phi(x, \mu_h))}{\sum_{h'=1}^{k} \pi_{h'} \exp(-d_\phi(x, \mu_{h'}))} \]

\[ d_{\beta \phi} = \beta d_\phi \]

Posterior probabilities are binarized when \( \beta \to \infty \)
Consider the update equation for the E-step

\[ p(h \mid x) = \frac{\pi_h \exp(-d_\phi(x, \mu_h))}{\sum_{h'=1}^k \pi_{h'} \exp(-d_\phi(x, \mu_{h'}))} \]

- \( d_{\beta \phi} = \beta d_\phi \)
- Posterior probabilities are binarized when \( \beta \to \infty \)
- Hey presto! A Bregman k-means algorithm
Outline

Introduction
- Standard EM algorithm
- Bregman Divergences
- Regular exponential families

An equivalence relationship
- A Legendre dual
- Exponential families to Bregman divergences
- Bregman divergences to regular exponential families

Clustering and Mixture Modeling
- Motivation
- The Bregman advantage
- Bregman k-means

Bregman Information
- Information-theoretic clustering

Nuisance parameters
Motivation

- We want to create $k$ disjoint partitions of a set $\mathcal{X}$ using an alphabet $\mathcal{M}$
- Quality of partitioning measured as loss of mutual information due to quantization
We have seen that Bregman hard clustering is equivalent to finding

$$\min_M \left( \sum_{h=1}^{k} \nu_h d_{\phi}(x_i, \mu_h) \right)$$
**Definition**

The optimal distortion-rate function of random variable $X$ for the Bregman divergence $d_{\phi}$ is called *Bregman information* and is given by

$$I_{\phi}(X) = \min_{s \in ri(S)} E_\nu[d_{\phi}(X, s)]$$
Bregman Information

Definition
The optimal distortion-rate function of random variable $X$ for the Bregman divergence $d_\phi$ is called \textit{Bregman information} and is given by

$$I_\phi(X) = \min_{s \in ri(S)} E_\nu[d_\phi(X, s)] = E_\nu[d_\phi(X, \mu)]$$
Clustering as loss of Bregman information

**Theorem**

Total Bregman information equals the sum of inter-cluster Bregman information and intra-cluster Bregman information, i.e.

$$I_{\phi}(X) = E_{\phi}[I_{\phi}(X_h)] + I_{\phi}(M)$$  \hspace{1cm} (6)
Outline

Introduction
  Standard EM algorithm
  Bregman Divergences
  Regular exponential families

An equivalence relationship
  A Legendre dual
  Exponential families to Bregman divergences
  Bregman divergences to regular exponential families

Clustering and Mixture Modeling
  Motivation
  The Bregman advantage
  Bregman k-means

Bregman Information
  Information-theoretic clustering

Nuisance parameters
Experiments

- Results from special cases of Bregman clustering well known
- What happens when Bregman divergence of algorithm and generative model differ?
- Metric for clustering - $I_{\phi}(X_{predicted}; X_{original})$
- Best performance seen for matching Bregman divergences
Outline

Introduction
  Standard EM algorithm
  Bregman Divergences
  Regular exponential families
An equivalence relationship
  A Legendre dual
  Exponential families to Bregman divergences
  Bregman divergences to regular exponential families
Clustering and Mixture Modeling
  Motivation
  The Bregman advantage
  Bregman k-means
Bregman Information
  Information-theoretic clustering
Nuisance parameters
Open questions

- Does there exist a larger class of Bregman divergences tractable to this analysis?
- Would it be interesting to analyze them?
- How would we select a specific Bregman divergence given domain knowledge?