Clustering with Bregman Divergences

Banerjee, Merugu, Dhillon and Ghosh, JMLR 2005

Nisheeth Srivastava

Dept of Computer Science
University of Minnesota

September 14, 2007

Outline

Iterative Expectation-Maximization

- Initialize $\{\theta_h, \pi_h\}_{h=1}^{k}$
- The expectation step

  for i $= 1$ to n **do**

  for h $= 1$ to k **do**

  $$p(h|\mathbf{x_i}) \leftarrow \frac{\pi_h \, p_{(\psi,\theta_h)}(\mathbf{x_i})}{\sum_{h'=1}^{k} \pi_{h'} \, p_{(\psi,\theta_{h'})}(\mathbf{x_i})}$$

  end for

  end for
- The Maximization step

  for h $= 1$ to k **do**

  $\pi_h \leftarrow \frac{1}{n} p(h \mid \mathbf{x_i})$

  $\theta_{\mathbf{h}} \leftarrow \text{argmax}_\theta \sum_{i=1}^{n} \log(p_{(\psi,\theta_h)}(\mathbf{x_i}))p(h \mid \mathbf{x_i})$

  end for
- **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
  $\theta_{\mathbf{h}} \leftarrow \text{argmax}_\theta \sum_{i=1}^{n} \log(p_{(\psi,\theta_h)}(\mathbf{x_i}))p(h \mid \mathbf{x_i})$

Problems with EM

- Max likelihood estimation
  - Local minima
  - M step computationally intensive
- Find a good replacement for
  $\theta_{\mathbf{h}} \leftarrow \text{argmax}_\theta \sum_{i=1}^{n} \log(p_{(\psi,\theta_h)}(\mathbf{x_i}))p(h \mid \mathbf{x_i})$

Outline

Bregman Divergence

- $d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \bigtriangledown \phi(\mathbf{y}) \rangle$

Bregman Divergence

- $d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \bigtriangledown \phi(\mathbf{y}) \rangle$
- Some properties

Bregman Divergence

- $d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \bigtriangledown \phi(\mathbf{y}) \rangle$
- Some properties
  - ϕ is strictly convex and is defined on convex set $\mathcal{S} \subseteq \mathbb{R}^d$

Bregman Divergence

- $d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \bigtriangledown \phi(\mathbf{y}) \rangle$
- Some properties
  - ϕ is strictly convex and is defined on convex set $\mathcal{S} \subseteq \mathbb{R}^d$
  - $d_\phi(\mathbf{x}, \mathbf{y}) \geq 0$, with equality only when $\mathbf{x} = \mathbf{y}$

Bregman Divergence

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

An example

$$
\begin{aligned}
d_\phi(\mathbf{x}, \mathbf{y}) &= \sum_{j=1}^{d} p_j \log_2 p_j - \sum_{j=1}^{d} q_j \log_2 q_j - \langle \mathbf{p} - \mathbf{q}, \nabla\phi(\mathbf{q}) \rangle \\
&= \sum_{j=1}^{d} p_j \log_2 p_j - \sum_{j=1}^{d} q_j \log_2 q_j - \\
&\quad \sum_{j=1}^{d} (p_j - q_j)(\log_2 q_j + \log_2 e) \\
&= \sum_{j=1}^{d} p_j \log_2 \left(\frac{p_j}{q_j} \right) - \log_2 e \sum_{j=1}^{d} (p_j - q_j) \\
&= \mathbf{KL}(\mathbf{p} \| \mathbf{q})
\end{aligned}
$$

Outline

Definition

- $\mathrm{p}_{(\psi,\theta)}(\mathbf{x}) = \exp(\langle \mathbf{x}, \theta \rangle - \psi(\theta))\, \mathrm{p}_0(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^d$

Definition

- $\mathrm{p}_{(\psi,\theta)}(\mathbf{x}) = \exp(\langle \mathbf{x}, \theta \rangle - \psi(\theta))\, \mathrm{p}_0(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^d$
- where,
  - $\mathbf{x}$ is a minimal natural statistic for the family
  - Parameter space Θ is open

Definition

- $\mathrm{p}_{(\psi,\theta)}(\mathbf{x}) = \exp(\langle \mathbf{x}, \theta \rangle - \psi(\theta))\, \mathrm{p}_0(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^d$
- where,
  - $\mathbf{x}$ is a minimal natural statistic for the family
  - Parameter space Θ is open
- Also,
  - Cumulant function (ψ) unique for a family
  - (Θ, ψ) 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

$$\mu = \mu(\theta) = \int_{\mathbb{R}^d} \mathbf{x} p_{(\psi,\theta)}(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

Outline

A legendary duel

Figure: The duel of Faerûn:Forgotten Realms

Legendre duality

Definition

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

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

Theorem

If (Θ, ψ) is a convex function of the Legendre type then

1. *The gradient function $\bigtriangledown\psi : \Theta \mapsto \Theta^\star$ is a one-to-one function from the open convex set Θ to the open convex set $\Theta^\star$*
2. *The gradient functions $\bigtriangledown\psi$ and $\bigtriangledown\psi^\star$ are continuous and $\bigtriangledown\psi^\star = (\bigtriangledown\psi)^{-1}$*

Expectation/natural parameter duality

$$\int p_{(\psi,\theta)}(\mathbf{x})\ d\mathbf{x} = 1$$

Differentiating with respect to θ

$$\int \frac{\partial}{\partial\theta} exp(\langle \mathbf{x}, \theta\rangle - \psi(\theta))\, p_0(\mathbf{x}) d\mathbf{x} = 0$$

Then

$$\mu = \mu(\theta) = \bigtriangledown\psi(\theta) \tag{1}$$

Define conjugate of ψ

$$\phi(\mu) = \sup_{\theta\in\mathbf{\Theta}}\{\langle \mu, \theta\rangle - \psi(\theta)\}$$

Dual space mapping

Θ and int(dom(ϕ)) will have the following mapping

$$\mu(\theta) = \bigtriangledown\psi(\theta) \quad \text{and} \quad \theta(\mu) = \bigtriangledown\phi(\mu) \tag{2}$$

Conjugate function can be expressed as

$$\phi(\mu) = \langle\theta(\mu), \mu\rangle - \psi(\theta(\mu)), \quad \forall\mu \in \mathit{int}(\mathit{dom}(\phi)) \tag{3}$$

Outline

A simple transformation

$$\begin{aligned}\langle \mathbf{x}, \theta \rangle - \psi(\theta) &= (\langle \mu, \theta \rangle - \psi(\theta)) + \langle \mathbf{x} - \mu, \theta \rangle \\ &= \phi(\mu) + \langle \mathbf{x} - \mu, \nabla \phi(\mu) \rangle \\ &= -d_\phi(\mathbf{x}, \mu) + \phi(\mathbf{x})\end{aligned}$$

where,

$\mathbf{x} \in dom(\phi),\ \theta \in \Theta$ and $\mu \in int(dom(\phi))$

A tricky alignment

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

where,

$$b_\phi(\mathbf{x}) = exp(\phi(\mathbf{x}))\, p_0(\mathbf{x})$$

A tricky alignment

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

where,

$$b_\phi(\mathbf{x}) = exp(\phi(\mathbf{x}))\, p_0(\mathbf{x})$$

- We don't know if I_ψ is identical with $\mathrm{dom}(\phi)$

A tricky alignment

Theorem

Let I_ψ be the set of instances that can be drawn following $p_{(\psi,\theta)}(\mathbf{x})$. Then $I_\psi \subseteq dom(\phi)$, where ϕ is the conjugate function of ψ

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

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

The main theorem

Theorem

Let $p_{(\psi,\theta)}$ be the probability density function of a regular exponential family distribution.Let ϕ be the conjugate function of ψ, so that (int(dom(ϕ)), ϕ) is the Legendre dual of (Θ, ψ).

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

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

The main theorem

Theorem

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

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

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

The main theorem

Theorem

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

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

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

Outline

Validity of the converse

- Regular exponential families $\mapsto$ Bregman divergence

Validity of the converse

- Regular exponential families $\mapsto$ Bregman divergence
- Bregman divergence $\mapsto$ 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 ν such that $f : \Theta \mapsto \mathcal{R}_{++}$ can be represented as

$$f(\theta) = \int_{\mathbf{x} \in \mathcal{R}^d} exp(\langle \mathbf{x}, \theta \rangle)\, d\nu(\mathbf{x}) \tag{4}$$

is that f is continuous and exponentially convex

Regular Bregman Divergence

Definition

Let ψ be a strictly convex function and ϕ be its conjugate. Then the Bregman divergence d_ϕ derived from ϕ is a regular Bregman divergence.

Validity of the converse

- Regular exponential families $\mapsto$ Bregman divergence

Validity of the converse

- Regular exponential families $\mapsto$ Bregman divergence
- Regular Bregman divergence $\mapsto$ Regular exponential family

Outline

Two sides of a coin

- Soft clustering $\leftrightarrow$ 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,

$$p(\mathbf{x} \mid \Gamma) =$$

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,

$$p(\mathbf{x} \mid \Gamma) = \sum_{h=1}^{k} \pi_h \; p_{(\psi,\theta_h)}(\mathbf{x})$$

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,

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

Outline

Iterative Expectation-Maximization

- Initialize $\{\theta_h, \pi_h\}_{h=1}^{k}$
- The expectation step

  for i $= 1$ to n **do**

    for h $= 1$ to k **do**

      $\mathrm{p}(\mathrm{h}|\mathbf{x_i}) \leftarrow \frac{\pi_h \, p_{(\psi,\theta_h)}(\mathbf{x_i})}{\sum_{h'=1}^{k} \pi_{h'} \, p_{(\psi,\theta_{h'})}(\mathbf{x_i})}$

    end for

  end for
- The Maximization step

  for h $= 1$ to k **do**

    $\pi_h \leftarrow \frac{1}{n} p(h \mid \mathbf{x_i})$

    $\theta_\mathbf{h} \leftarrow \mathrm{argmax}_\theta \sum_{i=1}^{n} \log(p_{(\psi,\theta_h)}(\mathbf{x_i}))p(h \mid \mathbf{x_i})$

  end for
- **until** convergence

Natural to unnatural

Maximization step

$$\theta_h = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \log(p_{(\psi,\theta)}(\mathbf{x}_i))\, p(h|\mathbf{x}_i)$$

is equivalent to

$$\begin{aligned}
\mu_h &= \operatorname*{argmax}_{\mu} \sum_{i=1}^{n} \log(b_\phi(\mathbf{x}_i)\, exp(-d_\phi(\mathbf{x}_i,\mu)))\, p(h|\mathbf{x}_i) \\
&= \operatorname*{argmax}_{\mu} \sum_{i=1}^{n} (\log(b_\phi(\mathbf{x}_i)) - d_\phi(\mathbf{x}_i,\mu)) p(h|\mathbf{x}_i) \\
&= \operatorname*{argmin}_{\mu} \sum_{i=1}^{n} d_\phi(\mathbf{x}_i,\mu) \frac{p(h \mid \mathbf{x}_i)}{\sum_{i'=1}^{n} p(h \mid \mathbf{x}_{i'})} \qquad (5)
\end{aligned}$$

A surprising result

Proposition

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

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

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

Bregman soft clustering

- Initialize $\{\theta_h, \pi_h\}_{h=1}^{k}$
- The expectation step

  for i $= 1$ to n **do**

    for h $= 1$ to k **do**

      $\mathrm{p}(\mathrm{h}|\mathbf{x_i}) \leftarrow \frac{\pi_h \, p_{(\psi,\theta_h)}(\mathbf{x_i})}{\sum_{h'=1}^{k} \pi_{h'} \, p_{(\psi,\theta_{h'})}(\mathbf{x_i})}$

    end for

  end for

- **until** convergence

Bregman soft clustering

- Initialize $\{\theta_h, \pi_h\}_{h=1}^{k}$
- The expectation step

  for i = 1 to n **do**

  for h = 1 to k **do**

  $$p(h|\mathbf{x_i}) \leftarrow \frac{\pi_h \, p_{(\psi,\theta_h)}(\mathbf{x_i})}{\sum_{h'=1}^{k} \pi_{h'} \, p_{(\psi,\theta_{h'})}(\mathbf{x_i})}$$

  end for

  end for
- The Maximization step

  for h = 1 to k **do**

  $$\pi_h \leftarrow \frac{1}{n} p(h \mid \mathbf{x_i})$$

  $$\mu_h \leftarrow \frac{\sum_{i=1}^{n} p(h|\mathbf{x}_i) \, \mathbf{x}_i}{\sum_{i=1}^{n} p(h|\mathbf{x}_i)}$$

  end for
- **until** convergence

Outline

Bregman hard clustering

- Consider the update equation for the E-step

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

Bregman hard clustering

- Consider the update equation for the E-step

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

- $\mathrm{d}_{\beta\phi} = \beta \mathrm{d}_\phi$

Bregman hard clustering

- Consider the update equation for the E-step

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

- $\mathrm{d}_{\beta\phi} = \beta \mathrm{d}_\phi$
- Posterior probabilities are binarized when $\beta \to \infty$

Bregman hard clustering

- Consider the update equation for the E-step

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

- $\mathrm{d}_{\beta\phi} = \beta \mathrm{d}_\phi$
- Posterior probabilities are binarized when $\beta \to \infty$
- Hey presto! A Bregman k-means algorithm

Outline

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

Bregman hard clustering

- We have seen that Bregman hard clustering is equivalent to finding

$$\min_{M} \left(\sum_{h=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{X}_h} \nu_i d_\phi(\mathbf{x}_i, \mu_h) \right)$$

Bregman Information

Definition

The optimal distortion-rate function of random variable X for the Bregman divergence d_ϕ is called *Bregman information* and is given by

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

Bregman Information

Definition

The optimal distortion-rate function of random variable X for the Bregman divergence d_ϕ is called *Bregman information* and is given by

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

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.

$$I_\phi(X) = E_\phi[I_\phi(X_h)] + I_\phi(M) \tag{6}$$

Outline

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

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 specfic Bregman divergence given domain knowledge?