Clustering with Bregman Divergences.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, Joydeep Ghosh

Xiaowei Song
Math 710

Oct 15, 2015
Instructor: Prof. Jacob Kogan

Outline, Banerjee et al. [2005]

Bregman Divergence Definition

Bregman, 1967; Censor and Zenios, 1998

Definition (Bregman Divergence)

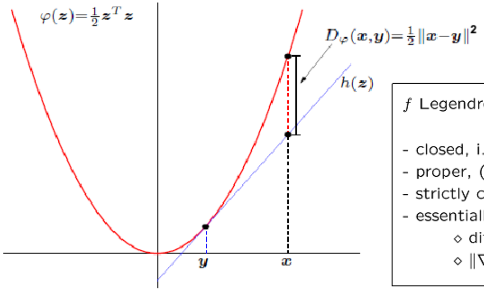
Let $\Phi : \mathcal{S} \mapsto \mathbb{R}$, $\mathcal{S} = dom(\Phi)$ be a strictly convex function defined on a convex set $\mathcal{S} \subseteq \mathbb{R}^d$ such that Φ is differentiable on $ri(\mathcal{S})$, assumed to be nonempty. The Bregman divergence $d_\Phi : \mathcal{S} \times ri(\mathcal{S}) \mapsto [0, \infty)$ is defined as:

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

,where $\nabla\Phi(y)$ represents the gradient vector of Φ evaluated at y.

Euclidean distance

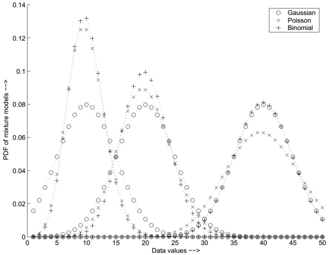
$\Phi(x) = \langle x, x \rangle$ strictly convex and differenable on $\mathbb{R}^d \Rightarrow$
$d_\Phi(x, y) = \langle x, x \rangle - \langle y, y \rangle - \langle x - y, 2y \rangle = \|x - y\|^2$
$d_\Phi(x, y) \geq 0$ as long as Φ convex
(http://mark.reid.name/blog/meet-the-bregman-divergences.html)

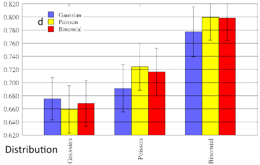


f Legendre function:

- closed, i.e. $\{x : f(x) \leq \alpha\}$ is closed
- proper, $(f > -\infty)$
- strictly convex
- essentially smooth
  - ⋄ differentiable
  - ⋄ $\|\nabla f(x_t)\| \to \infty$ when $x_t \to \mathrm{bd}(\mathrm{dom} f)$

Experiments about underlying distributions





Generative Model	$d_{Gaussian}$	$d_{Poisson}$	$d_{Binomial}$
Gaussian	0.675±0.032	0.659±0.036	0.668±0.035
Poisson	0.691±0.036	0.724±0.036	0.716±0.036
Binomial	0.777±0.038	0.799±0.0345	0.798±0.034

Each of 3 types' mixed density generated 300 points, were clustered 100 trials. Compared to ground-truth with NMI.

NMI is estimated based on “Evaluation of Clustering” (http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clu

KL-divergence

$\sum_{j=1}^{d} p_j = 1$, neg-entropy $\Phi(p) = \sum_{j=1}^{d} p_j log_2 p_j$ convex

$$
\begin{aligned}
d_\Phi(p,q) &= \sum_{j=1}^{d} p_j log_2 p_j - \sum_{j=1}^{d} q_j log_2 q_j - \langle p - q, \nabla\Phi(q)\rangle \\
&= \sum p_j log_2 p_j - \sum q_j log_2 q_j - \sum (p_j - q_j)(\log_2 q_j + \log_2 e) \\
&= \sum p_j \log_2\left(\frac{p_j}{q_j}\right) - (\log_2 e)\cdot \underbrace{\sum (p_j - q_j)}_{=0} \\
&= KL(p||q)
\end{aligned}
$$

for $f(p) = p\log_2 p$, $0 \le p \le 1$, $\frac{df}{dp} = \log_2 p + \log_2 e$, $\frac{d^2 f}{dp^2} = \frac{1}{p}\log_2 e > 0 \Rightarrow f(p)$ convex in $[0,1]$, thus $\sum f(p_j)$ convex in $0 \le p_j \le 1$

Itakura-Saito distance

If $F\left(e^{j\theta}\right)$ is the power spectrum of a signal $f(t)$, then the functional $\Phi(F) = -\frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left(F\left(e^{j\theta}\right)\right)d\theta$ is convex in F and corresponds to the neg-entropy rate of the signal assuming it was generated by a stationary Gaussian process.

$$
\begin{aligned}
d_{\Phi}(F,G) =& \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[-\log\left(F\left(e^{j\theta}\right)\right)+\log\left(G\left(e^{j\theta}\right)\right)\right. \\
& \left. -\left(F\left(e^{j\theta}\right)-G\left(e^{j\theta}\right)\right)\left(-\frac{1}{G\left(e^{j\theta}\right)}\right)\right]d\theta \\
=& \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(-\log\left(\frac{F(e^{j\theta})}{G(e^{j\theta})}\right)+\frac{F(e^{j\theta})}{G(e^{j\theta})}-1\right)d\theta
\end{aligned}
$$

Bregman divergences generated from convex functions

Domain	$\Phi(x)$	$d_\Phi(x,y)$	Divergence
$\mathbb{R}$	x^2	$(x-y)^2$	Squared loss
$\mathbb{R}^d$	$\|x\|^2$	$\|x-y\|^2$	Squared Euclidean distance
$\mathbb{R}^d$	x^TAx	$(x-y)^T A(x-y)$	Mahalanobis distance
$\mathbb{R}_+$	$x\log x$	$x\log\frac{x}{y}-(x-y)$	
d-Simplex	$\sum_{j=1}^d x_j\log_2 x_j$	$\sum_{j=1}^d x_j\log_2\frac{x_j}{y_j}-\log_2 e\times\left[\sum_{j=1}^d(x_j-y_j)\right]$	KL-divergence
$\mathbb{R}_+^d$	$\sum_{j=1}^d x_j\log x_j$	$\sum_{j=1}^d x_j\log\frac{x_j}{y_j}-\log_e e\times\left[\sum_{j=1}^d(x_j-y_j)\right]$	Generalized I-divergence
[0,1]	$x\log x+(1-x)\log(1-x)$	$x\log\frac{x}{y}+(1-x)\log\frac{1-x}{1-y}$	Logistic Loss
$\mathbb{R}_{++}$	$-\log x$	$\frac{x}{y}-\log\frac{x}{y}-1$	Itakura-Satio distance
$\mathbb{R}$	e^x	$e^x-e^y-(x-y)e^y$	

Function Name	$\varphi(x)$	$\operatorname{dom}\varphi$	$D_\varphi(x,y)$
Squared norm	$\frac{1}{2}x^2$	$(-\infty,+\infty)$	$\frac{1}{2}(x-y)^2$
Shannon entropy	$x\log x-x$	$[0,+\infty)$	$x\log\frac{x}{y}-x+y$
Bit entropy	$x\log x+(1-x)\log(1-x)$	$[0,1]$	$x\log\frac{x}{y}+(1-x)\log\frac{1-x}{1-y}$
Burg entropy	$-\log x$	$(0,+\infty)$	$\frac{x}{y}-\log\frac{x}{y}-1$
Hellinger	$-\sqrt{1-x^2}$	$[-1,1]$	$(1-xy)(1-y^2)^{-1/2}-(1-x^2)^{1/2}$
ℓ_p quasi-norm	$-x^p \quad (0<p<1)$	$[0,+\infty)$	$-x^p+p\,xy^{p-1}-(p-1)\,y^p$
ℓ_p norm	$\|x\|^p \quad (1<p<\infty)$	$(-\infty,+\infty)$	$\|x\|^p-p\,x\operatorname{sgn}y\,\|y\|^{p-1}+(p-1)\|y\|^p$
Exponential	e^x	$(-\infty,+\infty)$	$e^x-(x-y+1)e^y$

Hellinger:
$\varphi(\mathbf{x})=-\sqrt{1-\|\mathbf{x}\|^2}$

$D_\varphi(\mathbf{x},\mathbf{y})=$
$=\frac{1-\mathbf{x}^T\mathbf{y}}{\sqrt{1-\|\mathbf{y}\|^2}}-\sqrt{1-\|\mathbf{x}\|^2}$

Appendix A. Properties

1. Non-negativity. $d_\Phi(x,y) \geq 0, \forall x \in S,\ y \in ri(S)$, and equality holds IFF $x = y$. (Not a metric: not symmetric and triangle inequality not hold)
2. Convexity. d_Φ is always convex in the 1st argument, but not necessary convex in the 2nd argument. While, Squared Euclidean distance and KL-divergence are convex in both of their arguments.
3. Linearity. Bregman divergence is a linear operator, i.e.,

$$
\begin{aligned}
&\forall x \in S, y \in ri(S),\\
d_{\Phi_1+\Phi_2}(x,y) =& d_{\Phi_1}(x,y) + d_{\Phi_2}(x,y)\\
d_{c\Phi}(x,y) =& c d_\Phi(x,y),\ c \geq 0
\end{aligned}
$$

Appendix A. Properties

4. Equivalence classes. The Bregman divergences of functions that differ only in affine terms are identical, i.e., if$\Phi(x) = \Phi_0(x) + \langle b, x\rangle + c, b \in \mathbb{R}^d, c \in \mathbb{R}$, then $d_\Phi(x, y) = d_{\Phi_0}(x, y), \forall x \in S, y \in ri(S)$. Hence, the set of all strictly convex, differentiable functions on a convex set S can be partitioned into equivalence classes of the form

$$[\Phi_0] = \{\Phi \,|\, d_\Phi(x, y) = d_{\Phi_0}(x, y), \forall x \in S, y \in ri(S)\}$$

5. Linear separation.

$$\begin{aligned} d_\Phi(x, \mu_1) =& d_\Phi(x, \mu_2) \\ \Rightarrow \Phi(x) - \Phi(\mu_1) - \langle x - \mu_1, \nabla\Phi(\mu_1)\rangle =& \\ \Phi(x) - \Phi(\mu_2) - \langle x - \mu_2, \nabla\Phi(\mu_2)\rangle& \\ \Rightarrow \langle x, \nabla\Phi(\mu_2) - \nabla\Phi(\mu_1)\rangle =& \\ (\Phi(\mu_1) - \Phi(\mu_2)) - (\langle \mu_1, \nabla\Phi(\mu_1)\rangle - \langle \mu_2, \nabla\Phi(\mu_2)\rangle)& \end{aligned}$$

Appendix A. Properties

6 Dual divergences/Conjugate duality: let $\Psi(\theta) = \Phi^*(\theta)$ be the conjugate of $\Phi(u)$. Then $d_\Phi(\mu_1, \mu_2) = d_\Psi(\theta_2, \theta_1)$

$$\Psi(\theta) = \Phi^*(\theta) = \sup_u \underbrace{\left\{\theta^T u - \Phi(u)\right\}}_{g(\theta,u)}$$

Properties of conjugate function:

1). *let* $0 = \nabla_u g(\theta, u) = \theta - \nabla\Phi(u^*)$

2). Φ convex $\Rightarrow \Psi$ convex

3). Φ convex and closed $\Rightarrow (\Phi^*)^* = \Phi$

Proof of Conjugate duality

$$
\begin{aligned}
d_\Phi(u_1, u_2) =& \Phi(u_1) - \Phi(u_2) - (u_1 - u_2)^T \underbrace{\nabla\Phi(u_2)}_{\theta_2} \\
=& \Phi(u_1) - \Phi(u_2) - (u_1 - u_2)^T \theta_2 + \underbrace{u_1^T}_{\nabla\Psi(\theta_1)} \theta_1 - u_1^T \theta_1 \\
=& \Phi(u_1) - \Phi(u_2) - (\theta_2 - \theta_1)^T \nabla\Psi(\theta_1) + u_2^T \theta_2 - u_1^T \theta_1 \\
=& \left[\theta_2^T u_2 - \Phi(u_2)\right] - \left[\theta_1^T u_1 - \Phi(u_1)\right] - (\theta_2 - \theta_1)^T \nabla\Psi(\theta_1) \\
=& \Psi(\theta_2) - \Psi(\theta_1) - (\theta_2 - \theta_1)^T \nabla\Psi(\theta_1) \\
=& d_\Psi(\theta_2, \theta_1)
\end{aligned}
$$

[7]Relation to KL-divergence

Let $\mathcal{F}_\Psi$ be an exponential family with Ψ as the cumulant function.

$$KL\left(p_{(\Psi,\theta_1)} \middle\| p_{(\Psi,\theta_2)}\right) = d_\Psi(\theta_2, \theta_1) = d_\Phi(\mu_1, \mu_2)$$

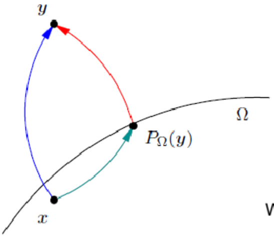
where $\mu_1, \mu_2$are the expectation parameters corresponding to θ_1, θ_2. Further, if $\Psi(0) = 0$, then $p_{(\Psi,0)}(x) = p_0(x)$ is itself a valid probability density and $KL\left(p_{(\Psi,\theta)} \middle\| p_{(\Psi,0)}\right) = \Phi(\mu)$, where $\mu = \nabla\Psi(\theta)$

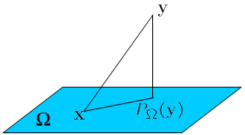
[8]Generalized Pythagoras theorem

Nearness in Bregman divergence:

the Bregman projection of y onto a convex set Ω.

$$P_\Omega(\mathbf{y}) = \arg\min_{\omega \in \Omega} D_\varphi(\omega, \mathbf{y})$$





When Ω is affine set, the Pythagoras theorem holds with equality.

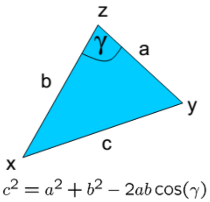
Generalized Pythagoras theorem:

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

Opposite to triangle inequality:

8

"Law of cosine"



Three point property generalizes the "law of cosine":
$D_\varphi(x,y) = D_\varphi(x,z) + D_\varphi(z,y) - (x-z)^T\,(\nabla\varphi(y) - \nabla\varphi(z))$
Euclidean special case:
$\|x-y\|^2 = \|x-z\|^2 + \|z-y\|^2 - 2\,(x-z)^T\,(y-z)$

Necessary & Sufficient conditions

A divergence measure $d : S \times ri(S) \mapsto [0, \infty)$ is a Bregman divergence IFF there exists $a \in ri(S)$ such that the function $\Phi_a(x) = d(x, a)$ satisfies the following conditions:

1. $\Phi_a(x)$ is strictly convex on S and differentiable on $ri(S)$
2. $d(x, y) = d_{\Phi_a}(x, y), \forall x \in S, y \in ri(S)$ where $d_{\Phi_a}$is the Bregman divergence associated with Φ_a

Proof of necessarity: any strictly convex, differentiable function Φ, the Bregman divergence evaluated with a fixed value for the 2nd argument differs from it only by a linear term, i.e.,
$\Phi_a(x) = d_\Phi(x, a) = \Phi(x) - \Phi(a) - \langle x - a, \nabla\Phi(a)\rangle =$
$\Phi(x) - \langle x, \nabla\Phi(a)\rangle - \Phi(a) + \langle a, \nabla\Phi(a)\rangle = \Phi(x) + \langle b, x\rangle + c$
where $b = -\nabla\Phi(a)$, $c = -\Phi(a) + \langle a, \nabla\Phi(a)\rangle$

Centroid

Proved in class by Prof. Kogan.

For the data points above $a_i, 1 \leq i \leq m$, we want to find one point closest to all data points,
define cost function: $f(x) = \sum_{i=1}^{m} |x - a_i|^2$, we want to get $\min_{x \in R^n} f(x)$, then use the found x to represent $a_i, 1 \leq i \leq m$
let $0 = \frac{df(x)}{dx} = \frac{d}{dx}\left[\sum (x - a_i)^2\right] = 2\sum(x - a_i) = 2\sum x - 2\sum a_i$
$\Rightarrow mx = \sum_{i=1}^{m} a_i \Rightarrow x = \frac{1}{m}\sum_{i=1}^{m} a_i$ which is the mean of all data points.

Proposition 1

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

$$\min_{s \in ri(\mathcal{S})} E_v[d_\Phi(X, s)]$$

has a unique minimizer given by $s^\dagger = \mu = E_v[X]$.
Note the minimization is with respect to 2nd argument, surprising since Bregman divergences are not necessarily convex in the 2nd argument.

Proposition 1 - Proof

The function we are trying to minimize is $J_\Phi(s) = E_v[d_\Phi(X, s)] = \sum_{i=1}^{n} v_i d_\Phi(x_i, s)$. Since $\mu = E_v[X] \in ri(\mathcal{S})$, the objective function is well-defined at μ. Now $\forall s \in ri(\mathcal{S})$,

$$\begin{aligned}
&J_\Phi(s) - J_\Phi(\mu) \\
=&\sum_{i=1}^{n} v_i d_\Phi(x_i, s) - \sum_{i=1}^{n} v_i d_\Phi(x_i, \mu) \\
=&\Phi(\mu) - \Phi(s) - \left\langle \sum_{i=1}^{n} v_i x_i - s,\ \nabla\Phi(s) \right\rangle + \left\langle \sum_{i=1}^{n} v_i x_i - \mu,\ \nabla\Phi(\mu) \right\rangle \\
=&\Phi(\mu) - \Phi(s) - \langle \mu - s,\ \nabla\Phi(s) \rangle \\
=&d_\Phi(\mu, s) \geq 0
\end{aligned}$$

with equality holds only when $s = \mu$

Bregman Information

Definition (Bregman Information)

Let X be a random variable that takes values in $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathcal{S}$ following a probability measure v. Let $\mu = E_v[X] = \sum_{i=1}^n v_i x_i \in ri(\mathcal{S})$ and let $d_\Phi : \mathcal{S} \times ri(\mathcal{S}) \mapsto [0, \infty)$ be a Bregman divergence. Then the Bregman Information of X in terms of d_Φ is defined as:
$I_\Phi(X) = E_v[d_\Phi(X, \mu)] = \sum_{i=1}^n v_i d_\Phi(x_i, \mu)$

Example 5. Variance:

Let $\mathcal{X} = \{x_i\}_{i=1}^n$ be a set in $\mathbb{R}^d$, and uniform measure $v_i = \frac{1}{n}$ over $\mathcal{X}$. The Bregman Information of X with sqared Euclidean distance as the Bregman divergence is given by:
$I_\Phi(X) = \sum_{i=1}^n v_i d_\Phi(x_i, \mu)$, which is sample variance

Bregman Information as Mutual information

Example 6. Mutual information:

By definition, the mutual information $I(U;\ V)$ between 2 discrete random variables U and V with joint distribution $\left\{\left\{p\left(u_i, v_j\right)\right\}_{i=1}^{n}\right\}_{j=1}^{m}$ is given by

$$\begin{aligned} I\left(U;\ V\right) &= \sum_{i=1}^{n}\sum_{j=1}^{m} p(u_i, v_j)\log\frac{p(u_i, v_j)}{p(u_i)p(v_j)} \\ &= \sum_{i=1}^{n} p(u_i)\sum_{j=1}^{m} p(\,v_j|\,u_i)\log\frac{p(\,v_j|\,u_i)}{p(v_j)} \\ &= \sum_{i=1}^{n} p(u_i) KL\left(p\left(\,V|\,u_i\right)\|\,p(V)\right) \end{aligned}$$

Consider RV Z_u taking values in the set of probability distributions $\mathcal{Z}_u = \{p\left(V|u_i\right)\}_{i=1}^{n}$ following the probability measure $\{v_i\}_{i=1}^{n} = \{p(u_i)\}_{i=1}^{n}$ over this set. The mean (distribution) of Z_u is given by:
$\mu = E_v\left[p(V|u)\right] = \sum_{i=1}^{n} p(u_i)p(V|u_i) = \sum_{i=1}^{n} p(u_i, V) = p(V)$
hence, $I(U,\ V) = \sum_{i=1}^{n} v_i d_\Phi(p(V|u_i), \mu) = I_\Phi(Z_u)$, similiarly, $I(U;\ V) = I_\Phi(Z_v)$

Jensen's Inequality and Bregman Information

Given any convex function Φ , for any random variable X ,
Jensen's inequality: $E[\Phi(X)] \geq \Phi(E[X])$

$$
\begin{aligned}
& E[\Phi(X)] - \Phi(E[X]) \\
=& E[\Phi(X)] - \Phi(E[X]) - \underbrace{E[\langle X - E[X], \nabla\Phi(E[X])\rangle]}_{0} \\
=& E[\Phi(X) - \Phi(E[X]) - \langle X - E[X], \nabla\Phi(E[X])\rangle] \\
=& E[d_\Phi(X, E(X))] \\
=& I_\Phi(X) \geq 0
\end{aligned}
$$

Clustering by Expected Bregman divergence

RV X takes values in $\mathcal{X} = \{x_i\}_{i=1}^n$ following prob measure v. When X has large Bregman information, it may not suffice to encode X using single representative since lower qunatization error may be desired.
Split the set $\mathcal{X}$ into k disjoint partitions $\{\mathcal{X}_h\}_{h=1}^k$, each with its own Bregman representative, RV M over the partition representatives as an appropriate quantization of X, which is $\mathcal{M} = \{\mu_h\}_{h=1}^k$, its probability as $\pi_h = \sum_{x_i \in \mathcal{X}_h} v_i$.
The quality of the quantization M can be measured by expected Bregman divergence between X and M, i.e., $E_{X,M}\left[d_\Phi(X, M)\right]$. Since M is a deterministic func of X , the expectation is actually over distribution of X ,

$$
\begin{aligned}
E_X\left[d_\Phi(X, M)\right] &= \sum_{h=1}^k \sum_{x_i \in \mathcal{X}_h} v_i d_\Phi(x_i, \mu_h) \\
&= \sum_{h=1}^k \pi_h \sum_{x_i \in \mathcal{X}_h} \frac{v_i}{\pi_h} d_\Phi(x_i, \mu_h) \\
&= E_\pi\left[I_\Phi\left(X_h\right)\right]
\end{aligned}
$$

Viewpoint as Information-theoretic clustering

In Infromation-theoretic clustering, the quality of partitioning is measured in terms of loss in mutual information resulting from the quantization of the original RV X, i.e., $I_\Phi(X) - I_\Phi(M)$.

Hard clustering problem is defined as finding a partitioning of $\mathcal{X}$, or equivalently, finding the random variable M , such that the loss in Bregman information due to quantization, $L_\Phi(M) = I_\Phi(X) - I_\Phi(M)$ is minimized.

Theorem (Information theoretic clustering)

Let X be a RV that takes values in $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathcal{S} \subseteq \mathbb{R}^d$ following positive probability measure v. Let $\{\mathcal{X}_h\}_{h=1}^k$ be a partitioning of $\mathcal{X}$ and let $\pi_h = \sum_{x_i \in \mathcal{X}_h} v_i$ be the induced measure π on the partitions. Let X_h be the RV that takes values in $\mathcal{X}_h$ following $\frac{v_i}{\pi_h}$ for $x_i \in \mathcal{X}_h$, $h = 1, \ldots, k$. Let $\mathcal{M} = \{\mu_h\}_{h=1}^k$ with $\mu_h \in ri\,(\mathcal{S})$ denote the set of representatives of $\{\mathcal{X}_h\}_{h=1}^k$, and M be a RV that takes values in $\mathcal{M}$ following π. then

$$L_\Phi(M) = I_\Phi(X) - I_\Phi(M) = E_\pi\left[I_\Phi(X_h)\right] = \sum_{h=1}^k \pi_h \sum_{x_i \in \mathcal{X}_h} \frac{v_i}{\pi_h} d_\Phi\left(x_i, \mu_h\right)$$

Information-theoretic clustering Proof

$I_\Phi(X)$

$$=\sum_{i=1}^{n} v_i d_\Phi(x_i,\mu) = \sum_{h=1}^{k}\sum_{x_i\in\mathcal{X}_h} v_i d_\Phi(x_i,\mu) = \sum_{h=1}^{k}\sum_{x_i\in\mathcal{X}_h} v_i \left\{\Phi(x_i)-\Phi(\mu)-\langle x_i-\mu,\nabla\Phi(\mu)\rangle\right\}$$

$$=\sum_{h=1}^{k}\sum_{x_i\in\mathcal{X}_h} v_i \left\{\Phi(x_i)-\Phi(\mu_h)-\langle x_i-\mu_h,\nabla\Phi(\mu_h)\rangle+\langle x_i-\mu_h,\nabla\Phi(\mu_h)\rangle\right.$$

$$\left.+\Phi(\mu_h)-\Phi(\mu)-\langle (x_i-\mu_h)+(\mu_h-\mu)\,,\nabla\Phi(\mu)\rangle\right\}$$

$$=\sum_{h=1}^{k}\sum_{x_i\in\mathcal{X}_h} v_i \left\{d_\Phi(x_i,\mu_h)+\langle x_i-\mu_h,\nabla\Phi(\mu_h)-\nabla\Phi(\mu)\rangle+d_\Phi(\mu_h,\mu)\right\}$$

$$=\sum_{h=1}^{k}\pi_h\sum_{x_i\in\mathcal{X}_h}\frac{v_i}{\pi_h}d_\Phi(x_i,\mu_h)+\sum_{h=1}^{k}\sum_{x_i\in\mathcal{X}_h} v_i d_\Phi(\mu_h,\mu)+\sum_{h=1}^{k}\pi_h\sum_{x_i\in\mathcal{X}_h}\frac{v_i}{\pi_h}\langle x_i-\mu_h,\nabla\Phi(\mu_h)-\nabla\Phi(\mu)\rangle$$

$$=\sum_{h=1}^{k}\pi_h I_\Phi(X_h)+\sum_{h=1}^{k}\pi_h d_\Phi(\mu_h,\mu)+\sum_{h=1}^{k}\pi_h\left\langle \underbrace{\sum_{x_i\in\mathcal{X}_h}\frac{v_i}{\pi_h}x_i}_{\mu_h}-\mu_h,\nabla\Phi(\mu_h)-\nabla\Phi(\mu)\right\rangle$$

$$=E_\pi\left[I_\Phi(X_h)\right]+I_\Phi(M)$$

Information-theoretic clustering interpretation

Within/Between cluster interpretation

- Total Bregman Information= $I_\Phi(X) = L_\Phi(M) + I_\Phi(M)$
- Within-cluster Bregman Information

$$= L_\Phi(M) = I_\Phi(X) - I_\Phi(M) = E_\pi\left[I_\Phi(X_h)\right] = \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h} v_i d_\Phi\left(x_i, \mu_h\right)$$

- Between-cluster Bregman Information= $I_\Phi(M)$

Using the theorem, Bregman clustering problem of minimizing the loss in Bregman information can be written as

$$\min_M \left(I_\Phi(X) - I_\Phi(M)\right) = \min_M \left(\sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h} v_i d_\Phi\left(x_i, \mu_h\right)\right)$$

Bregman Hard Clustering Algorithm

Input: Set $\mathcal{X} = \{x_i\}_{i=1}^{n} \subset \mathcal{S} \subseteq \mathbb{R}^d$, probability measure v over $\mathcal{X}$, Bregman divergence $d_\Phi : \mathcal{S} \times ri(\mathcal{S}) \mapsto \mathbb{R}$, number of clusters k.
Output: $\mathcal{M}^\dagger$, local minimizer of $L_\Phi(\mathcal{M}) = \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h} v_i d_\Phi(x_i, \mu_h)$ where $\mathcal{M} = \{\mu_h\}_{h=1}^{k}$, hard partitioning $\{\mathcal{X}_h\}_{h=1}^{k}$ of $\mathcal{X}$.
Method: Initialize $\{\mu_h\}_{h=1}^{k}$ with $\mu_h \in ri(\mathcal{S})$ (one possible initialization is to choose $\mu_h \in ri(\mathcal{S})$at random)
repeat
* The assignment Step
Set $\mathcal{X}_h \leftarrow \phi, 1 \leq h \leq k$
for i=1 to n do
$\mathcal{X}_h \leftarrow \mathcal{X}_h \cup \{x_i\}$
where $h = h^\dagger(x_i) = \arg\min_{h'} d_\Phi(x_i, \mu_{h'})$
endfor
* The Re-estimation Step
for $h = 1$ to k do
$\pi_h \leftarrow \sum_{x_i \in \mathcal{X}_h} v_i$
$\mu_h \leftarrow \frac{1}{\pi_h} \sum_{x_i \in \mathcal{X}_h} v_i x_i$
endfor
until convergence
return $\mathcal{M}^\dagger \leftarrow \{\mu_h\}_{h=1}^{k}$

Proof: Convergence and terminates in a finite steps at local optimal partition

The Bregman hard clustering algorithm monotonically decreases the loss function $\min_M (I_\Phi(X) - I_\Phi(M)) = \min_M \left(\sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h} v_i d_\Phi(x_i, \mu_h) \right)$.

Let $\left\{ \mathcal{X}_h^{(t)} \right\}_{h=1}^{k}$ be the partitioning of $\mathcal{X}$ after the t^{th} iteration and let $\mathcal{M}^{(t)} = \left\{ \mu_h^{(t)} \right\}_{h=1}^{k}$ be the corresponding set of cluster representatives. Then,

$$
\begin{aligned}
L_\Phi(M^{(t)}) &= \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h^{(t)}} v_i d_\Phi\left(x_i, \mu_h^{(t)}\right) \\
&\geq \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h^{(t)}} v_i d_\Phi\left(x_i, \mu_{h^\dagger(x_i)}^{(t)}\right) \\
&\geq \sum_{h=1}^{k} \sum_{x_i \in \mathcal{X}_h^{(t+1)}} v_i d_\Phi\left(x_i, \mu_h^{(t+1)}\right) = L_\Phi(M^{(t)})
\end{aligned}
$$

Properties of hard clustering

- **Exhaustiveness**: the algorithm works for all Bregman divergences and only for Bregman divergences since the arithmetic mean is the best predictor only for Bregman divergences.
- **Linear Separators:** The locus of points that are equidistant to 2 fixed points μ_1, μ_2 in terms of a Bregman divergence is given by $\mathcal{X} = \{x \mid d_\Phi(x, \mu_1) = d_\Phi(x, \mu_2)\}$, i.e., the set of points, $\{x \mid \langle x, \nabla\Phi(\mu_2) - \nabla\Phi(\mu_1)\rangle = (\Phi(\mu_1) - \Phi(\mu_2)) - (\langle \mu_1, \nabla\Phi(\mu_1)\rangle - \langle \mu_2, \nabla\Phi(\mu_2)\rangle)\}$
- **Scalability**: computational complexity of each iteration is linear in number of data points and number of desired cluster for all Bregman divergences.
- **Applicability to mixed data types**: One can choose different convex functions appropriate and meaningful for different subsets of the features. We can build a convex combination corresponding to Bregman divergence.

Exponential families
Sufficient Statistic

Consider a family $\mathcal{F}$ of probability densities on a measurable space $(\Omega, \mathcal{B})$ where $\mathcal{B}$ is a σ-algebra on the set Ω. Suppose every probability density, $p_\theta \in \mathcal{F}$, is parameterized by d real-valued variables $\theta = \{\theta_j\}_{j=1}^d$ so that
$\mathcal{F} = \left\{ p_\theta = f(\omega; \theta) \,\middle|\, \omega \in \mathcal{B}, \theta \in \Gamma \subseteq \mathbb{R}^d \right\}$.
Let $H : \mathcal{B} \mapsto \mathcal{G}$ transforms any RV $U : \mathcal{B} \mapsto \mathbb{R}$ to a RV $V : \mathcal{G} \mapsto \mathbb{R}$ with $V = H(U)$. Then given the probability density p_θ of U , H uniquely determines the probability density q_θ govering the RV V .

Definition (sufficient statistic)

If $\forall \omega \in \mathcal{B}$, $p_\theta(\omega)/q_\theta(\omega)$ exists and does not depend on θ, then H is called a sufficient statistic for the model $\mathcal{F}$.

Exponential families

Definition (exponential family, natural parameter)

If d-dimensional model $\mathcal{F} = \{p_\theta | \theta \in \Gamma\}$ can be expressed in terms of $(d+1)$real-valued linearly independent functions $\{C, H_1, \ldots, H_d\}$ on $\mathcal{B}$ and a function Ψ on Γ as $f(\omega;\theta) = \exp\left\{\sum_{j=1}^{d} \theta_j H_j(\omega) - \psi(\theta) + C(\omega)\right\}$, then $\mathcal{F}$ is called an exponential family, and θ is called its natural parameter.

If $\exists x \in \mathbb{R}^d$ such that $x_j = H_j(\omega)$, then density function $g(x;\theta) = \exp\left\{\sum_{j=1}^{d} \theta_j x_j - \psi(\theta) - \lambda(x)\right\}$ for a uniquely determined function $\lambda(x)$, is such that $f(\omega;\theta)/g(x;\theta)$ does not depend on θ. Thus x is sufficient statistic for the family.

Definition (exponential family, log-partition/cumulant function)

A multivariate parametric family $\mathcal{F}_\Psi$ of distribution $\left\{ p_{(\Psi,\theta)} \middle| \theta \in \Gamma \subseteq \mathbb{R}^d \right\}$ is called an exponential family if the probability density is of the form: $p_{(\Psi,\theta)} = \exp\left(\langle x, \theta\rangle - \psi(\theta) - \lambda(x)\right)$. The function $\psi(\theta)$ is known as log partition function or the cumulant function and it uniquely determines the exponential family $\mathcal{F}_\Psi$. Further, given $\mathcal{F}_\Psi$,ψ is uniquely determined up to a constant additive term. Amari [1995] showed Γ is a convex set in $\mathbb{R}^d$ and ψ is a strictly convex and differentiable function on $int(\Gamma)$.

Expectation parameters and Legendre duality

Consider a d-dimensional real RV X following an exponential family density $p_{(\psi,\theta)}$ specified by natural parameter $\theta \in \Gamma$. The expectation of X with respect to $p_{(\psi,\theta)}$, also called the expectation parameter, is given by: $\mu = \mu(\theta) = E_{p_{(\psi,\theta)}}[X] = \int_{\mathbb{R}^d} x p_{(\psi,\theta)}(x)dx$.
Amari [1995]showed that expectation and natural parameters have a one-one correspondence with each other and span spaces that exhibit a dual relationship.

Theorem (Rockafellar, 1970)

Let Ψ be a real-valued proper closed convex function with conjugate function Ψ^*.Let $\Theta = int(dom(\Psi))$ and $\Theta^* = int(dom(\Psi^*))$. If (Θ, Ψ) is a convex function of Legendre type, then

1. $(\Theta^*, \Psi^*)$is a convex function of Legendre type.
2. $(\Theta, \Psi)$and(Θ^*, Ψ^*) are Legendre duals of each other,
3. The gradient function $\nabla\Psi : \Theta \mapsto \Theta^*$ is a one-to-one function from the open convex set Θ onto the open convex set Θ^*
4. The gradient functions $\nabla\Psi, \nabla\Psi^*$ are continuous, and $\nabla\Psi^* = (\nabla\Psi)^{-1}$.

Expectation parameters and Legendre duality

Differentiating $1 = \int p_{(\psi,\theta)}(x)dx$ with respect to θ

$0 = \frac{\partial}{\partial\theta} \int \exp\left(\langle x, \theta\rangle - \psi(\theta) - \lambda(x)\right) dx =$

$\int \left(x - \nabla\psi(\theta)\right) p_{(\psi,\theta)}(x)dx$

$\Leftrightarrow \nabla\psi(\theta) \int p_{(\psi,\theta)}(x)dx = \int x p_{(\psi,\theta)}(x)dx$

$\Leftrightarrow \nabla\Psi(\theta) {=} \mu(\theta) {=} \mu$

Let Φ be defined as the conjugate of Ψ, i.e.,

$\Phi(\mu) = \sup_{\theta\in\Theta} \left\{\langle \mu, \theta\rangle - \Psi(\theta)\right\}$.

Then $\Phi = \Psi^*$ and $int(dom(\Phi)) = \Theta^*$, thus by Legendre transformation:

$\mu(\theta) = \nabla\Psi(\theta)$ and $\theta(\mu) = \nabla\Phi(\mu)$,

$\Phi(\mu) = \langle \theta(\mu), \theta\rangle - \Psi(\theta(\mu)), \forall \mu \in int\left(dom\left(\Phi\right)\right)$

Exponential Families and Bregman Divergences

$$\begin{aligned}\log\left(p_{(\psi,\theta)}(x)\right) &= \langle x,\theta\rangle - \psi(\theta) - \lambda(x)\\ &= [\langle \mu,\theta\rangle - \psi(\theta)] - \lambda(x) + \langle x-\mu,\theta\rangle\\ &= \Phi(\mu) + \langle x-\mu, \nabla\Phi(\mu)\rangle - \lambda(x)\\ &= [\Phi(\mu) + \langle x-\mu, \nabla\Phi(\mu)\rangle - \Phi(x)] + \Phi(x) - \lambda(x)\\ &= -d_\Phi\left(x,\mu(\theta)\right) + \Phi(x) - \lambda(x)\end{aligned}$$

Theorem (4. pdf expressed by Bregman Divergence)

Let $p_{(\psi,\theta)}$ be the pdf of a regular exponential family distribution. Let Φ be the conjugate function of Ψ so that $(int\,(dom\,(\Phi)),\Phi)$ is the Legendre dual of (Θ,Ψ). Let $\theta\in\Theta$ be the natural parameter and $\mu\in int\,(dom\,(\Phi))$ be the corresponding expectation parameter. Let d_Φ be the Bregman divergence derived from Φ. Then $p_{(\psi,\theta)}$ can be uniquely expressed as $p_{(\psi,\theta)}(x) = \exp\left(-d_\Phi\left(x,\mu\right)\right) b_\Phi(x), \forall x\in dom\,(\Phi)$, where $b_\Phi : dom\,(\Phi)\mapsto\mathbb{R}_+$ is a uniquely determined function.

Bijection with Regular Bregman Divergences

Theorem (Devinatz, 1955)

Let $\Theta \in \mathbb{R}^d$ be an open convex set. A necessary and sufficient condition that there exists a unique, bounded, non-negative measure v such that $f : \Theta \mapsto \mathbb{R}_{++}$ can be represented as $f(\theta) = \int_{x \in \mathbb{R}^d} \exp(\langle x, \theta \rangle)\, dv(x)$ is that f is continuous and exponentially convex.

Lemma 2. Let Ψ be the cumulant of an exponential family with base measure P_0 and natural parameter space $\Theta \in \mathbb{R}^d$. Then, if P_0 is concentrated on an affine subspaces of $\mathbb{R}^d$, then Ψ is not strictly convex.

Theorem (Bijection)

There is a bijection between regular exponential families and regular Bregman divergences.

Examples

Distribution	$p(x;\theta)$	μ	$\Phi(\mu)$	$d_\Phi(x,\mu)$
1-D Gaussian	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-a)^2}{2\sigma^2}\right]$	a	$\frac{1}{2\sigma^2}\mu^2$	$\frac{1}{2\sigma^2}(x-\mu)^2$
1-D Poisson	$\frac{1}{x!}\lambda^x\exp(-\lambda)$	λ	$\mu\log\mu-\mu$	$x\log\frac{x}{\mu}-(x-\mu)$
1-D Bernoulli	$q^x(1-q)^{1-x}$	q	$\mu\log\mu+(1-\mu)\log(1-\mu)$	$x\log\frac{x}{\mu}+(1-x)\log\frac{1-x}{1-\mu}$
1-D Binomial	$\binom{N}{x}q^x(1-q)^{N-x}$	Nq	$\mu\log\frac{\mu}{N}+(N-\mu)\log\frac{N-\mu}{N}$	$x\log\frac{x}{\mu}+(N-x)\log\frac{N-x}{N-\mu}$
1-D Exponential	$\lambda\exp(-\lambda x)$	$\frac{1}{\lambda}$	$-\log\mu-1$	$\frac{x}{\mu}-\log\frac{x}{\mu}-1$
d-D Sph. Gaussian	$\frac{1}{\sqrt{(2\pi\sigma^2)^d}}\exp\left[-\frac{\Vert x-a\Vert^2}{2\sigma^2}\right]$	a	$\frac{1}{2\sigma^2}\Vert\mu\Vert^2$	$\frac{1}{2\sigma^2}\Vert x-\mu\Vert^2$
d-D multinomial	$\frac{N!}{\prod_{j=1}^d x_j!}\prod_{j=1}^d q_j^{x_j}$	$[Nq_j]_{j=1}^{d-1}$	$\sum_{j=1}^d \mu_j\log\frac{\mu_j}{N}$	$\sum_{j=1}^d x_j\log\frac{x_j}{\mu_j}$

Distribution	θ	$\Psi(\theta)$	$dom(\Psi)$	$dom(\Phi)$	I_Ψ
1-D Gaussian	$\frac{a}{\sigma^2}$	$\frac{\sigma^2}{2}\theta^2$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
1-D Poisson	$\log\lambda$	$\exp\theta$	$\mathbb{R}$	$\mathbb{R}_+$	$\mathbb{N}$
1-D Bernoulli	$\log\frac{q}{1-q}$	$\log(1+\exp\theta)$	$\mathbb{R}$	$[0,1]$	$\{0,1\}$
1-D Binomial	$\log\frac{q}{1-q}$	$N\log(1+\exp\theta)$	$\mathbb{R}$	$[0,N]$	$\{0,1,\ldots,N\}$
1-D Exponential	$-\lambda$	$-\log(-\theta)$	$\mathbb{R}_{--}$	$\mathbb{R}_{++}$	$\mathbb{R}_{++}$
d-D Sph. Gaussian	$\frac{a}{\sigma^2}$	$\frac{\sigma^2}{2}\Vert\theta\Vert^2$	$\mathbb{R}^d$	$\mathbb{R}^d$	$\mathbb{R}^d$
d-D multinomial	$\left[\log\frac{q_j}{q_d}\right]_{j=1}^{d-1}$	$N\log\left(1+\sum_{j=1}^{d-1}\exp\theta_j\right)$	$\mathbb{R}^{d-1}$	$\left\{\mu\in\mathbb{R}_+^{d-1}, \vert\mu\vert\le N\right\}$	$\left\{x\in\mathbb{Z}_+^{d-1}, \vert x\vert\le N\right\}$

Example 9, spherical Gaussian distributions

$$
\begin{aligned}
p(x;a) &= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{1}{2\sigma^2}\|x-a\|^2\right) \\
&= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(\left\langle x, \frac{a}{\sigma^2}\right\rangle - \frac{\|a\|^2}{2\sigma^2} - \frac{\|x\|^2}{2\sigma^2}\right) \\
&= \exp\left(\langle x,\theta\rangle - \frac{\sigma^2}{2}\|\theta\|^2\right) \exp\left(-\frac{1}{2\sigma^2}\|x\|^2\right) \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \\
&= \exp\left(\langle x,\theta\rangle - \Psi(\theta)\right) p_0(x) \\
\therefore\ \mu &= \nabla\Psi(\theta) = \nabla\left(\frac{\sigma^2}{2}\|\theta\|^2\right) = \theta\sigma^2 = a \\
\therefore\ \Phi(\mu) &= \langle\mu,\theta\rangle - \Psi(\theta) = \left\langle \mu, \frac{\mu}{\sigma^2}\right\rangle - \frac{\sigma^2}{2}\|\theta\|^2 = \frac{\|\mu\|^2}{2\sigma^2} \\
\therefore\ d_\Phi(x,\mu) &= \Phi(x) - \Phi(\mu) - \langle x-\mu, \nabla\Phi(\mu)\rangle = \frac{\|x-\mu\|^2}{2\sigma^2} \\
b_\Phi(x) &= \exp(\Phi(x))\, p_0(x) = \exp\left(\frac{\|x\|^2}{2\sigma^2}\right) \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \\
p_{(\Psi,\theta)}(x) &= \exp\left(-d_\Phi(x,\mu)\right) b_\Phi(x)
\end{aligned}
$$

Soft clustering as a mixture density estimation

Definition (Bregman Soft clustering problem)

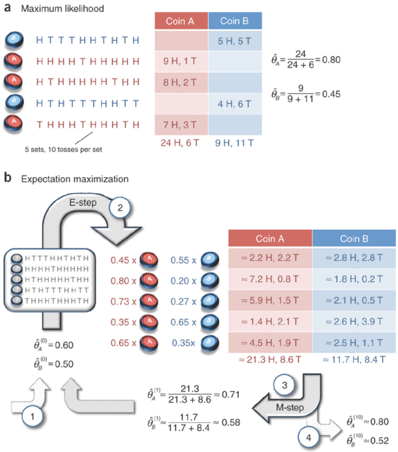
as that of of learning the maximum likelihood parameters $\Gamma = \{\theta_h, \pi_h\}_{h=1}^{k} \equiv \{\mu_h, \pi_h\}_{h=1}^{k}$ of a mixture model of the form

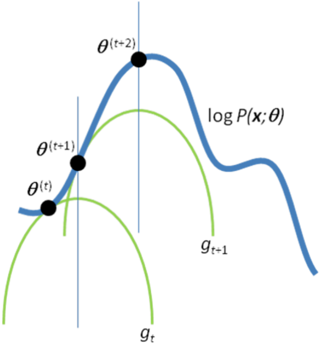
$$p(x|\Gamma) = \sum_{h=1}^{k} \pi_h p_{(\Psi,\theta_h)}(x) = \sum_{h=1}^{k} \pi_h \exp\left(-d_\Phi(x, \mu_h)\right) b_\Phi(x)$$

By assuming the mixture components from same family, it can be solved by EM algorithm.

EM example of coin flipping

Do and Batzoglou [2008]





EM for Mixture models based on Bregman Divergence

Algorithm 2 EM for Mixture Density Estimation [18]

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$, num. of clusters k.
Output: Θ^*, local maximizer of $L_{\mathcal{X}}(\Theta) = \prod_{i=1}^n (\sum_{h=1}^k \pi_h p_h(\mathbf{x}_i|\boldsymbol{\theta}_h))$ where $\Theta = \{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$, soft partitioning $\{\{p(h|\mathbf{x}_i)\}_{h=1}^k\}_{i=1}^n$.
Method:
Initialize $\{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$ with some $\boldsymbol{\theta}_h \in S$, $\pi_h \geq 0$, $\sum_{h=1}^k \pi_h = 1$
repeat
{The Expectation Step}
for $i = 1$ to n **do**
for $h = 1$ to k **do**
$p(h|\mathbf{x}_i) \leftarrow \frac{\pi_h p_h(\mathbf{x}_i|\boldsymbol{\theta}_h)}{\sum_{h'=1}^k \pi_{h'} p_{h'}(\mathbf{x}_i|\boldsymbol{\theta}_{h'})}$
end for
end for
{The Maximization Step}
for $h = 1$ to k **do**
$\pi_h \leftarrow \frac{1}{n}\sum_{i=1}^n p(h|\mathbf{x}_i)$
$\boldsymbol{\theta}_h \leftarrow \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{i=1}^n \log(p_h(\mathbf{x}_i|\boldsymbol{\theta}))p(h|\mathbf{x}_i)$
end for
until *convergence*
return $\Theta^* = \{\boldsymbol{\theta}_h, \pi_h\}_{h=1}^k$

Algorithm 3 Bregman Soft Clustering

Input: Set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subset S \subseteq \mathbb{R}^d$, Bregman divergence D_ϕ, num. of clusters k.
Output: Θ^*, local maximizer of $\prod_{i=1}^n (\sum_{h=1}^k \pi_h f_\phi(\mathbf{x}_i) \exp(-D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h)))$ where $\Theta = \{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$, soft partitioning $\{\{p(h|\mathbf{x}_i)\}_{h=1}^k\}_{i=1}^n$
Method:
Initialize $\{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$ with some $\boldsymbol{\mu}_h \in S, \pi_h \geq 0$, and $\sum_{h=}^k \pi_h = 1$
repeat
{The Expectation Step}
for $i = 1$ to n **do**
for $h = 1$ to k **do**
$p(h|\mathbf{x}_i) \leftarrow \frac{\pi_h \exp(-D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_h))}{\sum_{h'=1}^k \pi_{h'} \exp(-D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_{h'}))}$
end for
end for
{The Maximization Step}
for $h = 1$ to k **do**
$\pi_h \leftarrow \frac{1}{n}\sum_{i=1}^n p(h|\mathbf{x}_i)$
$\boldsymbol{\mu}_h \leftarrow \frac{\sum_{i=1}^n p(h|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^n p(h|\mathbf{x}_i)}$
end for
until *convergence*
return $\Theta^* = \{\boldsymbol{\mu}_h, \pi_h\}_{h=1}^k$

Extensions: Robustness

Geography faculty at the University of North Carolina like to point out that in 1986, those who graduated with a major in Geography had the highest average starting salaries in the class — $250,000. The punchline to this joke is that basketball legend, Michael Jordon, graduated from UNC with a major in Geography in 1986. In that particular dataset, Michael Jordan is clearly an outlier whose astronomical earnings skew the results and obscure the real market for geography majors. (Ref: http://www.forest2market.com/about/methodology/stumpage-price-database)

Definition (Robustness Check,Liu [2011])

Let $\bar{x}$ be the true centroid of set $X = \{x_1, \ldots, x_n\}$. When ϵ% (ϵ small) of outlier y is mixed into the set X, then the estimation of the centroid would be influnced by the outliers, and denote the estimation as $\hat{\bar{x}} = \bar{x} + \epsilon z(y)$, where the $z(y)$ is called the influnce function. For ordinary Bregman divergence, $z = y$,thus the breakdown point is 0%.

Extensions: Total Bregman divergence

Figure: Liu [2011] Ph.D Thesis page 16.

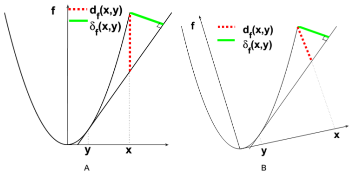


Figure 2-1. $d_f(x, y)$ (dotted line) is BD, $\delta_f(x, y)$ (bold line) is TBD, and the two arrows indicate the coordinate system. A $d_f(x, y)$ and $\delta_f(x, y)$ before rotating the coordinate system. B $d_f(x, y)$ and $\delta_f(x, y)$ after rotating the coordinate system.

Definition (Total Bregman divergence (TBD))

TBDδ associated with a real valued strictly convex and differentiable function f defined on a convex set X between points $x, y \in X$ is defined as,

$$\delta_f(x, y) = \frac{f(x) - f(y) - \langle x - y, \nabla f(y) \rangle}{\sqrt{1 + \|\nabla f(y)\|^2}}$$

Extensions: Symmetry

Definition (Symmetry Extension, Leonenko et al. [2008])

$$D_q(f,g) = \int_{\mathbb{R}^m} \left[g^q(x) + \frac{f^q(x)}{q-1} - \frac{q}{q-1} f(x) g^{q-1}(x) \right] dx$$

$$K_q(f,g) = \frac{1}{q} \left[D_q(f,g) + D_q(g,f) \right]$$

$$= \frac{1}{q-1} \int_{\mathbb{R}^m} \left[f(x) - g(x) \right] \left[f^{q-1}(x) - g^{q-1}(x) \right] dx$$

Reference

Shun-Ichi Amari. Information geometry of the EM and em algorithms for neural networks. *Neural networks*, 8(9): 1379–1408, 1995. URL http://www.sciencedirect.com/science/article/pii/0893608095000038. 00320 Cited by 0320.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005. URL http://dl.acm.org/citation.cfm?id=1194902. Cited by 0835.

Chuong B. Do and Serafim Batzoglou. What is the expectation maximization algorithm? *Nature Biotechnology*, 26 (8):897–899, August 2008. ISSN 1087-0156. doi: 10.1038/nbt1406. URL http://www.nature.com/nbt/journal/v26/n8/full/nbt1406.html. Cited by 0124.

Nikolai Leonenko, Luc Pronzato, and Vippal Savani. A class of Rényi information estimators for multidimensional densities. *The Annals of Statistics*, 36(5):2153–2182, October 2008. ISSN 0090-5364, 2168-8966. doi: 10.1214/07-AOS539. URL http://projecteuclid.org/euclid.aos/1223908088. 00141 Cited by 0141.

Meizhu Liu. *Total Bregman divergence, a robust divergence measure, and its applications*. PhD thesis, UNIVERSITY OF FLORIDA, 2011. URL http://gradworks.umi.com/35/14/3514960.html. Cited by 0003.

Acknowledgements

Thanks to:

- Prof. Jacob Kogan, Math 710