CS 395T	Large-Scale Data Mining	Fall 2004

Lecture 10
Information Theory, Clustering, and Bregman Divergences

Lecturer: Inderjit Dhillon Date: Oct. 28, Nov. 2

Keywords: *Clustering, information theory, relative entropy, mutual information, convexity, Bregman divergences*

1 Overview

In this lecture, we restrict our attention to "hard" clustering of the set $X = \{x_1, x_2, \ldots, x_n\}$ into k clusters, i.e.,

$$X = \bigcup_{j=1}^{k} \mathcal{X}_j, \quad \mathcal{X}_i \cap \mathcal{X}_j = \emptyset, \ \forall i \neq j$$

We develop an information-theoretic formulation of clustering and then show that this information-theoretic formulation and "usual" k-means with squared Euclidean distances may be regarded as special cases of clustering with Bregman divergences. The ideas in this lecture can be easily extended to develop agglomerative and divisive (repeated partitioning) algorithms.

2 Convexity

Convexity is a vast subject, with classics such as [?] entirely devoted to it. A modern book on convex optimization is by Boyd and Vandenberghe [?]. Convexity is the underlying theme in this lecture, so here we present some basic facts on convexity.

Definition 1 *A set S is convex if for every $x_1, x_2 \in S$ and $0 \leq \lambda \leq 1$, $\lambda x_1 + (1-\lambda)x_2 \in S$. The element $\lambda x_1 + (1-\lambda)x_2$, $0 \leq \lambda \leq 1$, is called a convex combination of x_1 and x_2.*

Definition 2 *A real-valued function $f : S \to \Re$ is said to be convex on S if for every $x_1, x_2 \in S$ and $0 \leq \lambda \leq 1$,*

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

f is strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Definition 3 *A function f is concave if $-f$ is convex.*

Obviously a convex function need not be differentiable ($f(x) = |x|$). However, it can be proved that a convex function on an open set S is continuous [?].

Theorem 4 *Suppose $f : S \to \Re$ is a differentiable function on any open convex set $S \subseteq \Re^d$. Then f is convex on S if and only if*

$$f(x) - f(y) \geq \nabla f(y)^T (x - y), \ \forall x, y \in S$$

where $\nabla f(y)$ is the n-dimensional gradient vector of f at y.

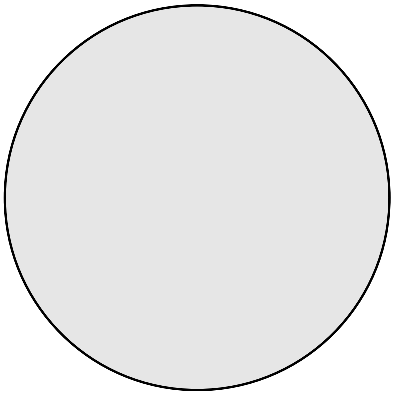

Figure 1: An example of a convex set in $\Re^2$

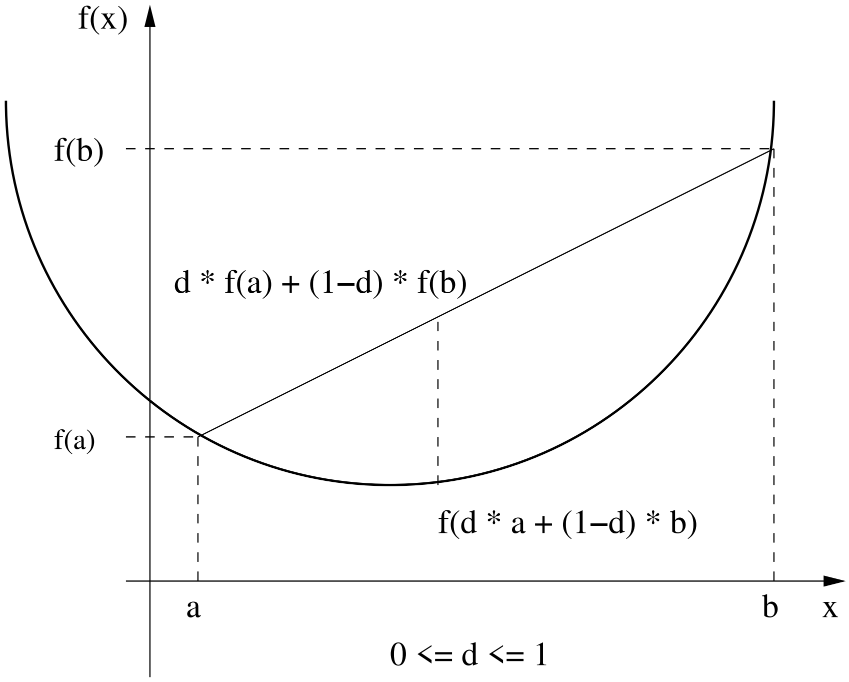

Figure 2: An example of a convex function

Proof: The proof is standard, for example see [**?**, Theorem 6.1.2]. ∎

When f is twice differentiable, there is an easy test for the convexity of f.

Theorem 5 *Suppose $f : S \to \Re$ is a twice-differentiable function on an open convex set $S \subseteq \Re^n$. Then f is convex on S if and only if, for each $y \in S$,*

$$x^T \nabla^2 f(y) x \geq 0, \;\; \forall x \in \Re^d$$

i.e., the $d \times d$ Hessian matrix $\nabla^2 f(y)$ is positive semi-definite. Similarly, strict convexity of f is equivalent to positive definiteness of the Hessian.

Proof: See [**?**, Theorems 6.3.1 and 6.3.2]. ∎

There are several ways of constructing convex functions from given convex functions, f_1 and f_2:

1. Any convex combination of f_1 and f_2, $\lambda f_1 + (1 - \lambda) f_2, \;\; 0 \leq \lambda \leq 1$, is a convex function.
2. Pointwise maximum of f_1 and f_2 (that have the same domain) is a convex function.

Examples of Convex Functions

1. $f : \Re^d \to \Re, \;\; f(x) = \|x\|_2^2$
2. $f : S \to \Re$, S is the d-dimensional unit simplex ($S = \{x \in \Re^d : x \geq 0$ and $\sum_{i=1}^n x(i) = 1\}$), $f(x) = \sum_{i=1}^d x(i) \log x(i)$.

3 Concepts from Information Theory

We now summarize some information-theoretic quantities before presenting our clustering formulation. For more details, see the excellent exposition in [**?**, Chapter 2].

Definition 6 *The entropy of a discrete random variable X, that takes values in the set $\mathcal{X}$ and has probability mass function $p(x) = Pr(X = x), \;\; x \in \mathcal{X}$, is defined by*

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

The convention is that $0 \log 0 = 0$, which is justified by continuity since $y \log y \to 0$ as $y \to 0$. $H(X)$ is often denoted by $H(p)$. Note that $H(X)$ can also be expressed as $\sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} = E_p\left[\log \frac{1}{p(x)}\right]$. Entropy can be interpreted as the measure of uncertainty of a random variable, or as the number of bits needed, on average, to describe/transmit/encode the random variable (we assume log is to the base 2).

Some Properties of Entropy

1. $H(X) \geq 0$
2. The uniform distribution, $u(x) = \frac{1}{|\mathcal{X}|}$, has the largest entropy, i.e.,

$$\forall p, \;\; H(p) \geq H(u)$$

3. $H(p)$ is a concave function of p.

Definition 7 *Relative entropy or Kullback-Liebler divergence between two probability mass functions $p(x)$ and $q(x)$ is defined as*

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \left[\log \frac{p(X)}{q(X)} \right]$$

The convention (again, based on continuity arguments) is that $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$ $(p \neq 0)$. Relative entropy is a measure of distance between two distributions, however, it is not a metric (neither symmetry nor the triangle inequality hold).

Some Properties of Relative Entropy

1. $D(p||q) \geq 0$ for all p, q.
2. In general, $D(p||q) \neq D(q||p)$ (not symmetric).
3. In general, it is **not** true that $D(p||r) \leq D(p||q) + D(q||r)$ (triangle inequality fails).
4. $D(p||q)$ is convex in the pair (p, q), i.e. if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then

$$D(\lambda p_1 + (1-\lambda) p_2 || \lambda q_1 + (1-\lambda) q_2) \leq \lambda D(p_1||q_1) + (1-\lambda) D(p_2||q_2), \quad \forall \lambda,\ 0 \leq \lambda \leq 1.$$

In many scenarios, we would like to measure the amount of information one random variable contains about another random variable. Mutual information is one such measure.

Definition 8 *Let X and Y be two random variables with joint probability mass function $p(x, y)$, and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information between X and Y is defined as*

$$I(X;Y) = D(p(X,Y)||p(X)p(Y)) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Since $I(X;Y)$ is the relative entropy between the joint distribution $p(X,Y)$ and the product distribution $p(X)p(Y)$, mutual information is always non-negative. Note that if X and Y are independent, then $p(x,y) = p(x)p(y)$ for all x and y, so $I(X;Y) = 0$.

Properties of Mutual Information

1. $I(X;Y) \geq 0$
2. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$, where $H(X,Y)$ is the joint entropy, while $H(X|Y)$ and $H(Y|X)$ are the conditional entropies, i.e.,

$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y)$$

$$H(X|Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x|y) = -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(x|y) \log p(x|y).$$

4 Information-Theoretic Clustering

Suppose the set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ is clustered to yield k clusters $\hat{x}_1, \ldots, \hat{x}_k$. Let $\hat{\mathcal{X}} = \{\hat{x}_1, \ldots, \hat{x}_k\}$. Let X and $\hat{X}$ be the random variables that take values in sets $\mathcal{X}$ and $\hat{\mathcal{X}}$ respectively. Let Y be another random variable that contains information about X, as measured by the mutual information $I(X;Y)$. Informally, a good clustering of $\mathcal{X}$ would be one that preserves "information" about Y. Formally, we can quantify the loss in mutual information upon clustering as $I(X;Y) - I(\hat{X};Y)$. We can write

$$I(X;Y) - I(\hat{X};Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} - \sum_{\hat{x}\in\hat{\mathcal{X}}}\sum_{y\in\mathcal{Y}} p(\hat{x},y)\log\frac{p(\hat{x},y)}{p(\hat{x})p(y)}$$

Since we are interested in hard clustering,

$$p(\hat{x}) = \sum_{x\in\hat{x}} p(x) \quad \text{and} \quad p(\hat{x},y) = \sum_{x\in\hat{x}} p(x,y) = \sum_{x\in\hat{x}} p(x)p(y|x),$$

$$\begin{aligned}
I(X;Y) - I(\hat{X};Y) &= \sum_{y\in\mathcal{Y}}\sum_{\hat{x}\in\hat{\mathcal{X}}}\sum_{x\in\hat{x}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} - \sum_{y\in\mathcal{Y}}\sum_{\hat{x}\in\hat{\mathcal{X}}}\sum_{x\in\hat{x}} p(x,y)\log\frac{p(\hat{x},y)}{p(\hat{x})p(y)} \\
&= \sum_{y\in\mathcal{Y}}\sum_{\hat{x}\in\hat{\mathcal{X}}}\sum_{x\in\hat{x}} p(x,y)\log\frac{p(x,y)}{p(\hat{x},y)\frac{p(x)}{p(\hat{x})}} \\
&= \sum_{\hat{x}\in\hat{\mathcal{X}}}\sum_{x\in\hat{x}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{p(y|x)}{p(y|\hat{x})} \\
&= \sum_{\hat{x}\in\hat{\mathcal{X}}}\sum_{x\in\hat{x}} p(x) D(p(Y|x)||p(Y|\hat{x})) \qquad (1)
\end{aligned}$$

where $p(Y|\hat{x}) = \sum_{x\in\hat{x}} p(x|\hat{x})p(Y|x)$. The expression (1) is suggestive of how to find a clustering $\hat{X}$ that tries to minimize the loss in mutual information, $I(X;Y) - I(\hat{X};Y)$. A possible algorithm alternates between (i) computing $p(Y|\hat{x})$ given the clustering $\hat{X}$, and (ii) computing a new clustering $\hat{X}$ given $p(Y|\hat{x})$. The algorithm is reminiscent of k-means with squared Euclidean distances, except that it uses relative entropy as the distance measure. The algorithm is given below.

Information-Theoretic Clustering

1. Start with some partitioning of the data into k clusters: $\hat{x}_1^{(0)}, \hat{x}_2^{(0)}, \ldots, \hat{x}_k^{(0)}$. Set iteration count $t \leftarrow 0$.

2. Calculate the mean of each cluster based on the current data partitioning:

$$\begin{aligned}
p(\hat{x}^{(t)}) &= \sum_{x\in\hat{x}^{(t)}} p(x) \\
p(Y|\hat{x}^{(t)}) &= \frac{1}{p(\hat{x}^{(t)})}\sum_{x\in\hat{x}^{(t)}} p(x)p(Y|x).
\end{aligned}$$

3. Update partitions based on the new cluster means as follows. For each $p(Y|x)$, find its new cluster index as

$$i^*(x) = \arg\min_i D(p(Y|x)||p(Y|\hat{x}^{(t)})),$$

resulting ties arbitrarily (when $p(Y|x)$ is equidistant to more than one $p(Y|\hat{x}^{(t)})$). Then compute the new clusters as

$$\hat{x}^{(t+1)} = \{x : i^*(x) = i\}.$$

4. If not "converged", $t \leftarrow t+1$ and go to step 2 (the convergence criterion can be $J^{(t)} - J^{(t+1)} \leq 10^{-3}$), where $J = I(X;Y) - I(\hat{X};Y)$, see (1). Otherwise, stop.

Recall that k-means with squared Euclidean distance monotonically decreases the corresponding objective function value. A crucial property in proving this result was that the mean vector of a collection of vectors minimizes the sum-squared deviation, i.e. $m = \frac{1}{n}\sum_{i=1}^n x_i = \arg\min_z \sum_{i=1}^n \|x_i - z\|^2$. Can a similar claim be made about the above algorithm, i.e., does the algorithm monotonically decrease the objective function in (1)? The answer is YES, and relies on the following lemma:

Lemma 9 *Given probability vectors* $x_1, x_2, \ldots, x_n \in \Re^d$ ($x_i \geq 0,\ \|x_i\|_1 = 1$) *and weights* $\pi_1, \pi_2, \ldots, \pi_n$ ($\pi_i \geq 0,\ \sum_{i=1}^n \pi_i = 1$)*, then*

$$\sum_{i=1}^n \pi_i D(x_i||m) \leq \sum_{i=1}^n \pi_i D(x_i||z)$$

where z ranges over all probability vectors and $m = \sum_{i=1}^n \pi_i x_i$.

Proof: Note that $\|m\|_1 = 1$, and

$$\begin{aligned}
\sum_{i=1}^n \pi_i \left(D(x_i||z) - D(x_i||m_i)\right) &= \sum_{i=1}^n \pi_i \sum_{j=1}^d x_i(j) \log \frac{x_i(j)}{z(j)} \cdot \frac{m(j)}{x_i(j)} \\
&= \sum_{j=1}^d \sum_{i=1}^n \pi_i x_i(j) \log \frac{m(j)}{z(j)} \\
&= \sum_{j=1}^d m(j) \log \frac{m(j)}{z(j)} \\
&= D(m||z) \geq 0 \quad \blacksquare
\end{aligned}$$

The above information-theoretic clustering algorithm has been applied to the problem of clustering words for the purpose of text classification in [?]. An information-theoretic co-clustering algorithm, which simultaneously clusters X and Y is given in [?].

5 Clustering with Bregman Divergences

The above section described a k-means algorithm that uses relative entropy as a similarity measure, and we saw the analogy with squared Euclidean distance. This section unifies the treatment by showing that both are special cases of Bregman divergences.

Definition 10 *Let ϕ be a strictly convex, differentiable function $\phi : S \rightarrow \Re$, where $S \subseteq \Re^d$ is a convex set. The Bregman divergence corresponding to ϕ, $D_\phi : S \times int(S) \rightarrow \Re$ is defined as*

$$D_\phi(x, y) = \phi(x) - \phi(y) - (x - y)^T \nabla\phi(y)$$

where $int(S)$ denotes the interior of S and is the maximal open set contained in S.

Due to the convexity of ϕ, it is easily seen that $D_\phi(x, y) \geq 0$ for all x, y (see Theorem 4 in Section 2). However, it need not be symmetric and generally the triangle inequality fails.

Examples

1. $S = \Re^d,\ \phi(x) = \|x\|_2^2,\ D_\phi(x,y) = \|x-y\|_2^2$
2. $S = \{x : x \geq 0 \text{ and } \|x\|_1 = 1\},\ \phi(x) = \sum_i x(i)\log x(i),\ D_\phi(x,y) = D(x\|y)$, i.e., the Begman divergence is the same as the relative entropy in this case.
3. $S = \Re_{++},\ \phi(x) = -\log x,\ D_\phi(x,y) = \frac{x}{y} - \log\frac{x}{y} - 1$, which is the Itakura-Saito distance that is widely used in signal processing.

Examples 1 and 2 above show that the squared Euclidean distance and relative entropy are special cases of Begman divergences. Given a set of points $x_1, \ldots, x_n \in S$, the following lemma shows that the mean vector is the optimal "Bregman representative."

Lemma 11 *Let D_ϕ be a Bregman divergence, $D_\phi : S \times int(S) \to \Re$. Let $x_1, \ldots, x_n \in S$ and let $\pi_1, \ldots, \pi_n$ be non-negative weights ($\sum_{i=1}^n \pi_i = 1$). Then*

$$\sum_{i=1}^n \pi_i D_\phi(x_i, m) \leq \sum_{i=1}^n \pi_i D_\phi(x_i, z), \quad \forall z \in int(S)$$

where $m = \sum_{i=1}^n \pi_i x_i$.

Thus a k-means algorithm for Bregman divergences can be obtained and can be shown to monotonically decrease a global loss function. For more details, including the connection between Bregman divergences and exponential families, see [?].