21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

On the Relation Between K-means and PLSA

Arghya Roy Chaudhuri
Dept. of EE (SSA)
Indian Institute of Science
Bangalore-560012
India
arghya.iisc@gmail.com

M. Narasimha Murty
Prof. Dept. of CSA
Indian Institute of Science
Bangalore-560012
India
mnm@csa.iisc.ernet.in

Abstract

Non-negative matrix factorization [5] (NMF) is a well known tool for unsupervised machine learning. It can be viewed as a generalization of the K-means clustering, Expectation Maximization based clustering and aspect modeling by Probabilistic Latent Semantic Analysis (PLSA). Specifically PLSA is related to NMF with KL-divergence objective function. Further it is shown that K-means clustering is a special case of NMF with matrix L2 norm based error function. In this paper our objective is to analyze the relation between K-means clustering and PLSA by examining the KL-divergence function and matrix L2 norm based error function.

1 Comparison between K-means and PLSA

There are several differences and common properties between K-means and PLSA. The basic difference is that K-means(hard or soft) is solved by minimizing the squared euclidean error and PLSA is solved by minimizing the KL-divergence. On the other hand K-means is a center based algorithm while PLSA is a generative model, it tries to solve the problem using the "aspect model" [3]. But there are several important properties shared by K-means and PLSA. K-means also uses Expectation Maximization and suffers from local minima problem. Another good commonality is that both of them can be generalized to NMF.

One thing that is clear from the theory is that both the squared euclidean norm and KL-divergence try to minimize entry wise error. But it is also clear that in the first case precision of accuracy for all the entries are equally treated where as the second one treats each entry with different weights or importance. As our objective is to set up a relation between these two let's have a closer look at the KL-divergence.

2 K-means and NMF

Let us have a closer look at bridging between two apparently different tools. As NMF acts as a platform for generalization of different tools we will look at different tools in the light of NMF. There are already brilliant works proving relation between NMF and K-means clustering[2], here we provide a brief discussion about this.

Let us consider a collection of m dimensional non-negative data as $A = \{\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_n}\} \in {\Re_+}^{m \times n}$, we can consider it as collection of n documents representing a word-document association between word i and document j; $\forall i = 1, ..., m$ and $\forall j = 1, ..., n$. The NMF factorizes X into two non negative matrices $F = \{\mathbf{f_1}, \mathbf{f_2}, ..., \mathbf{f_k}\} \in {\Re_+}^{m \times k}$ and $G = \{\mathbf{g_1}, \mathbf{g_2}, ..., \mathbf{g_k}\} \in {\Re_+}^{n \times k}$ such that :

$$A \approx FG^T \tag{1}$$

where F and G are indicator vectors for row cluster and column cluster respectively. With out loss of generality the j^{th} element (i.e the j^{th} column) of F looks like :

$$\mathbf{f_j} = (0, \cdots, 0, \overbrace{1, \cdots, 1}^{n_k}, 0, \cdots, 0)^T / n_k^{\frac{1}{2}} \tag{2}$$

Here $n_j = |C_j| =$ Cardinality of the j^{th} row cluster. For column cluster G analogously we can define each of its elements .

Now let $s(R_k, C_l) = \sum_{i \in R_k} \sum_{j \in C_l} b_{ij}$ be the total similarity value between row cluster R_k and column cluster C_l. K-means maximizes within cluster similarities $s(R_k, C_k)$,

$$\max_{\substack{F^TF=I;\\ G^TG=I;\\ F,G\geq 0}} J_1 = \sum_k \frac{s(R_k, C_k)}{(|R_k||C_k|)^{1/2}} = Tr(F^TAG)$$

$$\begin{aligned} \Rightarrow \min_{\substack{F^TF=I;\\ G^TG=I;\\ F,G\geq 0}} J_1 &= \min_{\substack{F^TF=I;\\ G^TG=I;\\ F,G\geq 0}} -2Tr(F^TAG) \\ &= \min_{\substack{F^TF=I;\\ G^TG=I;\\ F,G\geq 0}} ||A||^2 - 2Tr(F^TAG) \\ &\quad +Tr(F^TFG^TG) \\ &= \min_{\substack{F^TF=I;\\ G^TG=I;\\ F,G\geq 0}} ||A - FG^T||^2 \end{aligned} \tag{3}$$

Very clearly orthogonality is playing an important role. When $F^TF = I$ and $G^TG = I$ columns of F and G are orthogonal among themselves. This leads to hard K-means. But if we relax the orthogonality as $F^TF \approx I$ and $G^TG \approx I$ we will land-up on NMF which can be viewed as soft clustering. There[2] is a detailed discussion about how a column normalized A matrix (say Y) can be approximately factored into two non-negative matrices[2]

$$C = (\mathbf{c_1}, \cdots, \mathbf{c_k}), B = (\mathbf{b_1}, \cdots, \mathbf{b_k})$$

such that

$$Y \approx CB^T \tag{4}$$

with the normalization

$$\sum_{j=1}^{m} C_{jk} = \sum_{r=1}^{k} B_{ir} = 1$$

and how it can be approximated by hard and fuzzy K-means clustering by solving the following:

$$\min_{C,B\geq 0} J_{NMF} = ||Y - CB^T||^2 \tag{5}$$

It is shown[2] that this reduces to hard K-means when $B^TB = I$ and fuzzy K-means if we relax this condition.

3 A Closer Look At KL-divergence

For any $p, q \in \Re_+^m$ with $p = (p_1, p_2, \cdots, p_m)^T$ and $q = (q_1, q_2, \cdots, q_m)^T$ the generalized Kullback-Leibler divergence[1] [4] is defined as:

$$D_{KL}(p||q) = \sum_{i=1}^{m} (p_i \log \frac{pi}{q_i} - p_i + q_i) \tag{6}$$

Let $\rho_{max} > \rho_{min} > 0$ and $p, q \in [\rho_{min}, \rho_{max}]^m$. Consider the strictly convex, C^2 (i.e. continuously twice differentiable) function $\phi_{KL} : \Re^m \mapsto \Re$ defined as:

$$\phi_{KL}(t) = \sum_{i=1}^{m} t_i \log(t_i) - t_i \tag{7}$$

Now come to the following lemma

Lemma 1. *$D_{KL}(p||q)$ can be shown as the tail of the first-order Taylor expansion of $\phi_{KL}(p)$ at q.*

Proof. The k^{th}-order Taylor expansion of $\phi_{KL}(p)$ at q is given by

$$\phi_{KL}(p) = \sum_{\alpha=0}^{k} \frac{D^\alpha \phi(q)}{\alpha!}(p-q)^\alpha + \sum_{\beta=k+1} R_\beta(p)(p-q)^\beta \tag{8}$$

where the remainder R_β is given by

$$|R_\beta(p)| = \frac{|\beta|}{\beta!} \int_0^1 (1-t)^{|\beta|-1} D^\beta \phi(q+t(p-q))dt \tag{9}$$

Now simply put $k = 1$ and hence $\beta = 2$ we will get the value of the remainder same as R.H.S of equation (6) □

By Lagrange remainder form, $\exists \nu = (\nu_1, \nu_2, \cdots, \nu_m)^T \in \Re_+^m$ with $max(p_i, q_i) \geq \nu_i \geq min(p_i, q_i)$ such that

$$\begin{aligned} D_{KL}(p||q) &= \sum_{i,j=1}^{m} \frac{\partial^2}{2\partial t_i \partial t_j} \phi_{KL}(\nu_i) d_{ij} \\ &= (p-q)^T \nabla^2 \phi_{KL}(\nu)(p-q) \end{aligned} \tag{10}$$

where $d_{ij} = (p_i - q_i)(p_j - q_j)$ and the Hessian is given by

$$\nabla^2 \phi_{KL}(\nu) = \begin{bmatrix} \frac{1}{\nu_1} & & & \\ & \frac{1}{\nu_2} & & \\ & & \ddots & \\ & & & \frac{1}{\nu_m} \end{bmatrix}_{m \times m}$$

4 The Relation Between K-means and PLSA

Equation (10) is very much similar to squared Mahalanobis distance(see Mahalanobis [1936]). But the basic difference is KL-divergence is not a metric and hence the diagonal entries in the Hessian matrix will get changed to some other values for $D_{KL}(q||p)$. But one interesting thing to notice is that we can bound $D_{KL}(p||q)$ from both sides with Mahalanobis distance as follows: The squared Mahalanobis distance between two points $x, y \in \Re^m$ w.r.t. matrix Σ as

$$D_\Sigma(x,y) = (x-y)^T \Sigma (x-y) \tag{11}$$

It is obvious that $\exists \rho_{max}, \rho_{min} \in \Re$ s.t $\rho_{max} \geq \nu_i \geq \rho_{min}$ and therefore $\frac{1}{\rho_{max}} \leq \frac{1}{\nu_i} \leq \frac{1}{\rho_{min}}$.

Now if we define:

$$\Sigma_{\rho_{max}} = \begin{bmatrix} \frac{1}{\rho_{max}} & & & \\ & \frac{1}{\rho_{max}} & & \\ & & \ddots & \\ & & & \frac{1}{\rho_{max}} \end{bmatrix}_{m \times m},$$

$$\Sigma_{\rho_{min}} = \begin{bmatrix} \frac{1}{\rho_{min}} & & & \\ & \frac{1}{\rho_{min}} & & \\ & & \ddots & \\ & & & \frac{1}{\rho_{min}} \end{bmatrix}_{m \times m}$$

then using equation (10) and (11) we have

$$D_{\Sigma_{\rho_{max}}}(p,q) \leq D_{KL}(p||q) \leq D_{\Sigma_{\rho_{min}}}(p,q) \tag{12}$$

implies that

$$\begin{aligned} (p-q)^T \Sigma_{\rho_{max}} (p-q) &\leq D_{KL}(p||q) \\ &\leq (p-q)^T \Sigma_{\rho_{min}} (p-q) \end{aligned} \tag{13}$$

$$\Rightarrow \frac{1}{\rho_{max}} D_I(p,q) \leq D_{KL}(p||q) \leq \frac{1}{\rho_{min}} D_I(p,q) \tag{14}$$

where

$$\begin{aligned} D_I(p,q) &= (p-q)^T I (p-q) \\ &\Rightarrow \text{Squared Euclidean Distance} \end{aligned}$$

Now recall the objective function for K-means (equation (3)) and taking $FG^T = M$ we can write

$$V_A = (A_{11}, A_{21}, \cdots, A_{m1}, A_{12} \cdots, A_{mn})$$
$$V_M = (M_{11}, M_{21}, \cdots, M_{m1}, M_{12} \cdots, M_{mn})$$

as two vectors of length mn each, then from equation (3) and (5) it follows that solution to K-means problem is obtained by solving the following:

$$\begin{aligned} \min_{M \geq 0} J_1 &= ||A - M||^2 \\ &= (V_A - V_M)^T I (V_A - V_M) \\ &= D_I(V_A, V_M) \end{aligned} \tag{15}$$

Hence using the relation (14) we see : $\exists \eta_{max}, \eta_{min} \in \Re^+$ s.t $V_A, V_M \in [\eta_{min}, \eta_{max}]^{mn}$ and

$$\begin{aligned} \frac{1}{\eta_{max}} D_I(V_A, V_M) &\leq D_{KL}(V_A||V_M) \\ &\leq \frac{1}{\eta_{min}} D_I(V_A, V_M) \end{aligned} \tag{16}$$

So from the above relation we see that squared weighted Euclidean is a bad approximation of KL divergence as it does not take any special care about each of the terms. Obviously one can take that special care about each term using some weights. But then it is not guaranteed that those weights would be appropriate w.r.t. KL-divergence for the given context. Only when $V_A = V_M$ then both are minimized but that is also almost impractical due to presence of local minima and numerical errors in computation. Hence in general NMF with KL-divergence error function (which is equivalent to PLSA) outperforms the NMF with squared Euclidean error function.

Now another important point to relate K-means and PLSA is that K-means works on a metricspace where as PLSA works on a non-metric space. But if we investigate further we will see that there is a metric corresponding to KL-divergence so that it gets reduced on reducing the KL-divergence. Firstly we see that it can be shown[6] that the expected Mutual Information between two discrete random variables X, Y can be equivalently expressed as:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right) \tag{17}$$

where $p(x,y)$ is the joint probability distribution function of X and Y, and $p_1(x)$ and $p_2(y)$ are the marginal probability distribution functions of X and Y respectively. Mutual Information[6] can also be written as

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned} \tag{18}$$

where $H(X)$ and $H(Y)$ are the marginal entropies, $H(X|Y)$ and $H(Y|X)$ are the conditional entropies, and $H(X,Y)$ is the joint entropy of X and Y. This relation straightway leads to the following:

$$\begin{aligned} H(X,Y) - I(X,Y) &= H(X|Y) + H(Y|X) \\ &= d(X,Y)(\text{say}) \end{aligned} \tag{19}$$

Clearly this $d(X,Y)$ is non-negative, symmetric and follows triangular inequality and hence it is a metric. Dividing both sides by $H(X,Y)$ (as $d(X,Y) \leq H(X,Y)$) we get:

$$\frac{H(X,Y) - I(X,Y)}{H(X,Y)} = \frac{d(X,Y)}{H(X,Y)} = U(\text{say})$$

$$\Rightarrow U = 1 - \frac{I(X,Y)}{H(X,Y)} \tag{20}$$

which is effectively Jaccard distance.

Again we know that

$$\max\{H(X), H(Y)\} \leq H(X,Y) \leq H(X) + H(Y) \tag{21}$$

When the KL-divergence between $P(X)$ and $P(Y)$ is minimized we can assume that $P(Y) \approx P(X)$. Therefore $H(Y) \approx H(X)$ and hence from equation (18) we get that

$$I(X,Y) \approx H(X)$$

But it is also true that

$$H(X) = I(X,X) \geq I(X,Y) \tag{22}$$

Therefore reducing the KL-divergence between $P(X)$ and $P(Y)$ maximizes the numerator of the second term in equation (20) and minimizes the denominator simultaneously which leads to minimization of U in equation (20).

5 Conclusion and Future Work

In this paper, for the first time we formally analysed the relation between PLSA and K-means. Based on the analysis we come to the conclusion that NMF with KL-divergence as the error function and K-means are mutually related and K-means is a relaxed version of this. In future we plan to take advantage of this generalization to make PLSA to work faster. We are also interested to work on generalization among different machine learning tools for topic based clustering.

References

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, http://www.worldcat.org/isbn/0471062596, 1991.

[2] X. Ding, Chris. He and H. D. Simon. On the equivalence of nonnegative matrix factorization and k-means – spectral clustering. *Proc. SIAM Int'l Conf. Data Mining*, pages 606 – 610, May 2005.

[3] T. Hofmann. Probabilistic latent semantic indexing. *Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*, August 1999.

[4] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 1951.

[5] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788 – 791, Oct 1995.

[6] M. Meila. Comparing clusterings by the variation of information. *Learning Theory and Kernel Machines*, pages 173–187, August 2003.