Turn in your code along with your results in hard copy. Note that the assignment is due at 5PM IN TAY 137.

1. (5 points) Recall that the $k$-means algorithm can be generalized beyond the squared Euclidean distance to use any Bregman divergence. Implement the $k$-means algorithm when the divergence used is the KL divergence. The clustering objective function can be expressed as

$$
\min_{\pi_1, \ldots, \pi_K} \sum_{c=1}^{K} \sum_{x \in \pi_c} KL(x, m_c),
$$

where $\pi_c$ denotes the $c$-th cluster. Assume that each instance vector $x$ is strictly positive with $L_1$ norm equal to one (and hence each mean $m_c$ also has $L_1$ norm equal to one). Note that the KL divergence can be defined as

$$
KL(x, y) = \sum_i x_i \log(\frac{x_i}{y_i}),
$$

where $x$ and $y$ are vectors having $L_1$ norm equal to one.

2. (5 points) Implement the single-link agglomerative clustering algorithm using KL divergence. Your implementation should take in the number of agglomerated clusters $K$ as an input argument and run until there are only $K$ agglomerated clusters.

3. (5 points) Download the iris data set from http://www.cs.utexas.edu/~wtang/cs391d/iris.tar.gz. Run the above algorithms on this data set. Compare your clustering results to the “true” classes by computing the confusion matrix. Give the theoretical running time for both the agglomerative clustering algorithm as well as $k$-means. Also, compute the observed running time of a sample run over the data set of each algorithm. Note that you need to normalize each instance in iris data to have $L_1$ norm equal to one before applying your implemented algorithms.

4. (5 points) Let $d_\phi(x, y)$ denote the Bregman divergence between $x$ and $y$. We define

$$
S_T = \sum_x d_\phi(x, m), \quad S_B = \sum_{c=1}^{K} N_c d_\phi(m_c, m), \quad \text{and} \quad S_W = \sum_{c=1}^{K} \sum_{x \in \pi_c} d_\phi(x, m_c),
$$

where $m = \frac{1}{N} \sum_x x$ and $m_c = \frac{1}{N_c} \sum_{x \in \pi_c} x$ ($N$ is the total number of instances and $N_c$ is the number of instances in cluster $\pi_c$).

Show that $S_T = S_B + S_W$. 