CAP6610: Machine Learning
Homework 6

1. **Principal Component Analysis and Dictionary Learning:** Open the file clockwork-angels.jpg using a function like imread in octave (or MATLAB). Take just the first dimension (R of RGB) of the 1713x3448x3 multi-dimensional array. Create 1000 random 16x16 patches that index into the 1713x3448 array (using a function such as rand(1000,1) and scale it appropriately to get patches that stay within the image dimensions after first converting from integer to double precision).

   (a) Run PCA and dictionary learning [OMP, KSVD (in MATLAB) or anything else suitable discovered by googling]. When running dictionary learning, try and estimate 350 vectors in the dictionary (leading to an overcomplete representation). Write your own code to compare the results of PCA with DL and report your findings.

   (b) With the dictionary fixed (and henceforth referred to as W), write a majorization-minimization algorithm based on the development in class to minimize the objective function

   $$E_{\mathrm{DL}}\left(z\right) = \sum_{i=1}^{N} \|x_i - W z_i\|_2^2 + \lambda \sum_{i=1}^{N} \|z_i\|_1$$

   by first converting it into the majorized form

   $$E_{\mathrm{DL-maj}}(z, u) = \sum_{i=1}^{N} \|x_i - W z_i\|_2^2 + \lambda \sum_{i=1}^{N} \sum_{k=1}^{L} \frac{\left(z_{ik}^2 + u_{ik}^2\right)}{2u_{ik}}$$

   where $u_{ik} \geq \epsilon$, $\forall ik$. Implement an alternating algorithm that cycles between least-squares updates of z and updates of u until convergence. Compare your results by running SPAMS or equivalent. Report the ℓ_2 distance between the z vectors obtained by your method and those of the competing method.

2. **Clustering using mixture models:** Implement Gaussian mixture model (GMM) and K-Means clustering algorithms on the uploaded shape dataset and a dataset of your own choosing. In the mixture model, show that each update (including π and σ) lowers (or keeps fixed) the "complete data" negative log-likelihood objective function

   $$E_{\mathrm{cmp}}(r, \mu, \sigma, \pi) = -\sum_{i=1}^{N} \sum_{k=1}^{N} \left[r_{ik} \log\left\{\pi_k p(x_i|\mu_k, \sigma)\right\} + r_{ik} \log r_{ik}\right]$$

   to be minimized subject to the constraints $\pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1$, $r_{ik} \geq 0$ and $\sum_{k=1}^{K} r_{ik} = 1$. We did not update π in our code and are leaving it up to you. Here the density function

   $$p(x|\mu_k, \sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}\, \sigma^D} \exp\left\{-\frac{\|x - \mu_k\|_2^2}{2\sigma^2}\right\}.$$

   Execute a cluster quality criterion to decide on the relative merits of the GMM (if any) against K-Means clustering.