Homework 6

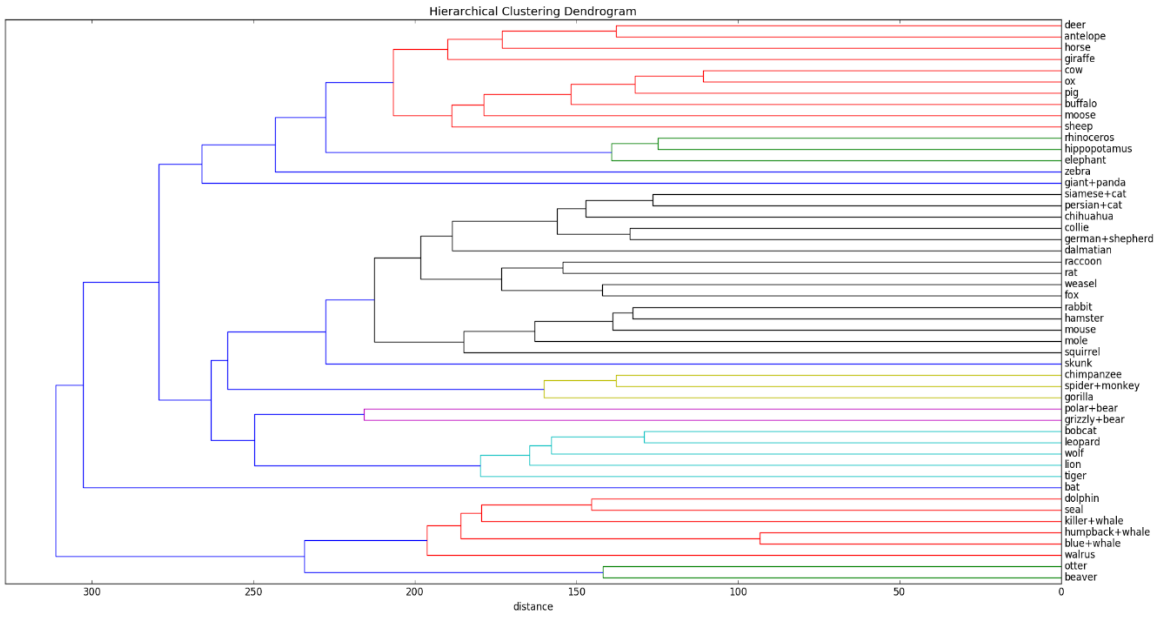
Problem 1:

The following is the list of k-means clusters:

Cluster No.	Animals
0.	Dalmatian, Persian cat, German shepherd, Siamese cat, Chihuahua, collie
1.	Moose, ox, sheep, buffalo, giant panda, pig, cow
2.	Spider monkey, gorilla, chimpanzee
3.	Killer whale, blue whale, humpback whale, seal, otter, walrus, dolphin
4.	Polar bear, grizzly bear
5.	Tiger, leopard, fox, wolf, bobcat, lion
6.	hippopotamus, elephant, rhinoceros
7.	beaver, skunk, mole, hamster, squirrel, rabbit, rat, weasel, mouse, raccoon
8.	antelope, horse, giraffe, zebra, deer
9.	bat

The clusters makes sense since animals with similar features are in the same cluster.

Dendron Diagram:



The hierarchical clusters makes sense since similar animals are grouped first.

Problem 2:

a. Consider a cluster C. Let n be the number of points x in the cluster, so that:
$x \in \{x^{(1)}, x^{(2)}, \dots x^{(n)}\}$
To find the optimal μ, we need to minimize the function with respect to μ:
$f = \sum_{j=1}^{n} ||x^{(j)} - \mu||^2 = \sum_{j=1}^{n} \sum_{i=1}^{p} (x_i^{(j)} - \mu_i)^2$
Taking partial derivative of f with respect to μ_i (ie ∂f_i) we get,
$\partial f_i = \sum_{j=1}^{n} 2 * (x_i^{(j)} - \mu_i) * (-1) = 0$
$\sum_{j=1}^{n} (x_i^{(j)} - \mu_i) = 0$
Hence $\mu_i = \frac{\sum_{j=1}^{n} x_i^{(j)}}{n}$ for all values of i
Thus we can write $\mu = \frac{\sum_{j=1}^{n} x^{(j)}}{n}$ ie, μ is the mean of points in C.
Hence proved.

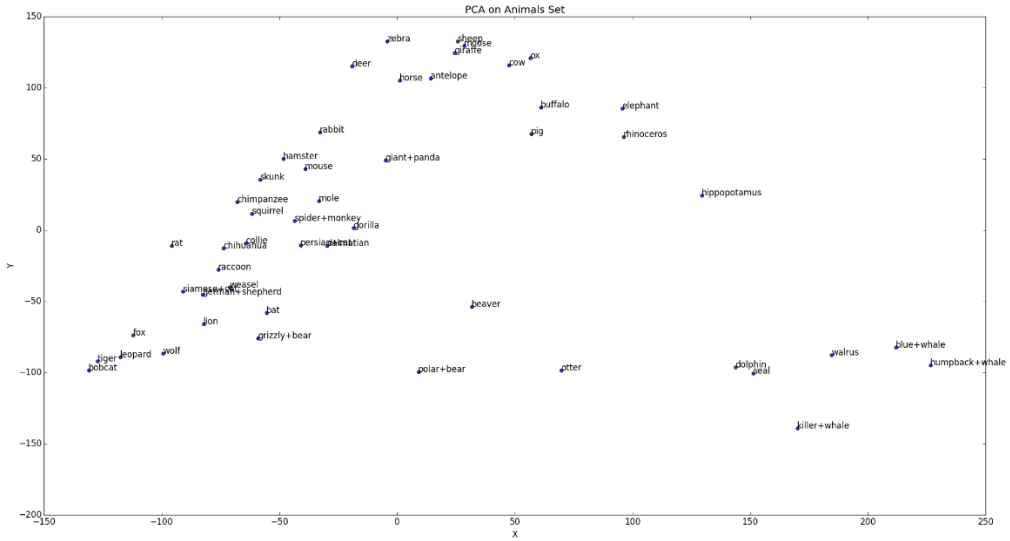
b. Consider three points in a cluster $x^{(1)} = 0$, $x^{(2)} = 4$, $x^{(3)} = 5$ in one dimension. According to k means the optimal center of the cluster should be $\mu = \frac{0+4+5}{3} = 3$. If l1 distance is chosen, then $f(x) = \sum_{j=1}^{3} |x^{(j)} - \mu| = |0 - 3| + |4 - 3| + |5 - 3| = 6$. But if $\mu = 4$ then f(x) = 5.
Hence l1 distance cannot be used for k means.
For (R1, l1), the optimal center can be characterized as the median of points in the cluster.

Problem 3:

a. The optimal k-means solution is : [-9, 0, 9]
b. Let the initialization of the centers be [0, 8, 10]. In this case, the final answer will have centers as [-6, 8, 10] and clusters as [{-10, -8, 0}, {8}, {10}] which is not the optimal solution.

Problem 4:

Visualization of Animal set in 2d by performing PCA:



The graph makes sense since similar animals are grouped nearby.

Problem 5:

a. The following are the dimensions of the given matrices in the form (rows, columns)
   - Dim(U) = (p, 2)
   - Dim(U^T) = (2, p)
   - Dim(UU^T) = (p, p)
   - Dim($u_1u_1^T$) = (p, p)

b. Projection 1 and 3 are the same. They represent the projection of x in a 2d subspace defined by (u1, u2)
   Projection 2 and 4 are the same. The represent projection of x on 2 directions defined by u1 and u2. It shows the projection of x as a p dimensional vector.