# CSE250B Homework 6

Qiao Zhang

February 29, 2016

## 1 Experiments with clustering

### 1.1 a

### 1.2 b

The list of k-means cluster is as follows:
0: ['beaver', 'skunk', 'mole', 'hamster', 'squirrel', 'rabbit', 'rat', 'mouse', 'raccoon']
1: ['killer+whale', 'blue+whale', 'humpback+whale', 'seal', 'otter', 'walrus', 'dolphin']
2: ['fox', 'wolf', 'weasel']
3: ['antelope', 'horse', 'moose', 'giraffe', 'zebra', 'deer']
4: ['hippopotamus', 'elephant', 'ox', 'sheep', 'rhinoceros', 'buffalo', 'giant+panda', 'pig', 'cow']
5: ['dalmatian', 'persian+cat', 'german+shepherd', 'siamese+cat', 'chihuahua', 'collie']
6: ['tiger', 'leopard', 'bobcat', 'lion']
7: ['spider+monkey', 'gorilla', 'chimpanzee']
8: ['bat']
9: ['grizzly+bear', 'polar+bear']
The result makes some sense but could be better. Usually k-means clustering would result in a big cluster containing items difficult to classify. But the problem is not that serious here.

### 1.3 c

The hierarchical clustering seems rather sensible to me, where animals similar or in the same order/family are clustered first.

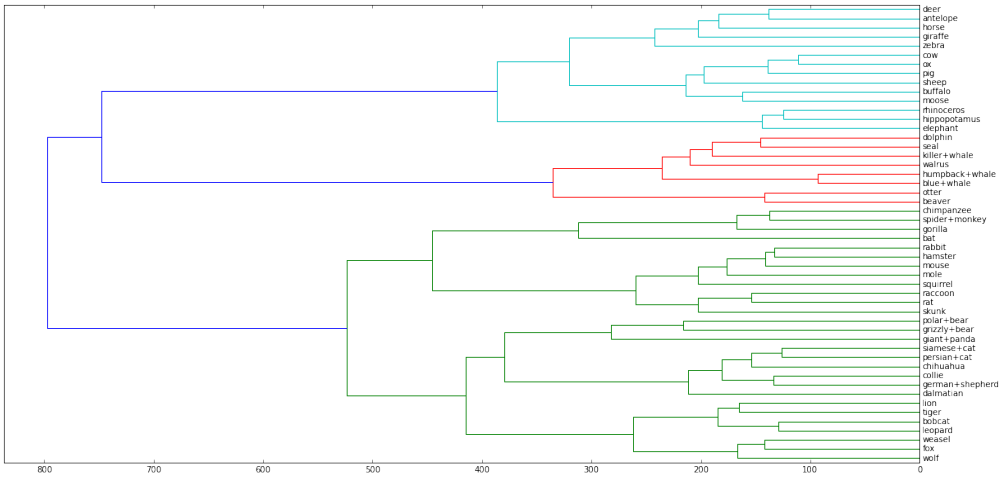


Figure 1: dendrogram, ward method

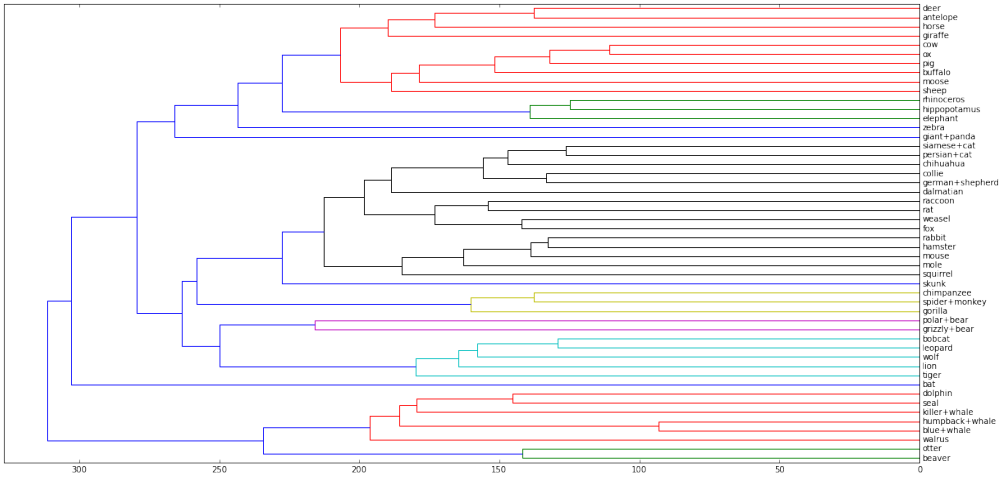


Figure 2: dendrogram, average method

# 2 Placement of the cluster center

## 2.1 a

Let $d = \sum_{x \in C} \|x - \mu\|^2$
then $\frac{\partial d}{\partial \mu_i} = \sum_{x \in C}(-2x_i) + 2\mu_i$
Let $\frac{\partial d}{\partial \mu_i} = 0$ to get the optimum
then $\mu_i = \frac{\sum x_i}{\|C\|}$, $\mu = \frac{\sum x}{\|C\|}$
Thus $\mu = mean(C)$ is the optimal center.

## 2.2 b

I can give a counterexample to show that this is not true. If the data points are 3, 5 and 10. The mean is 6 while the optimal solution is 5.
The optimal center location in (R1, l1) case should be point whose total distance to all points are minimized rather than total distance square.

# 3 A bad case for k-means

## 3.1 a

The optimal solution is -9, 0, 9.

## 3.2 b

If the initialization of the centers are 7, 8 and 10, the group formation will be (-10, -8, 0)(8)(10). The updated centers would be -6, 8 and 10, which is convergent. This final answer is sub-optimal.

# 4 An experiment with PCA

The methodology is that (1) calculate orthogonal basis based on eigenvectors (2) select the directions with larger variance and leave out the ones with small eigenvalues (3) project the data into new sub space and retrieve new feature. That is to say, we calculated 85 basis but had only the 2 most principal left.

The results are as follows. This embedding is not perfect but reasonable enough. Generally speaking, similar animals have shorter distance to each other. We can see some obvious reasonable clusters. For example, dolphin, seal and whale all lie on the right bottom corner while fox, leopard, tiger lie on the left bottom corner. But there still remains some problems in that a bunch of animals in the middle of the image are not clearly and well divided, which might result from dimension reduction.

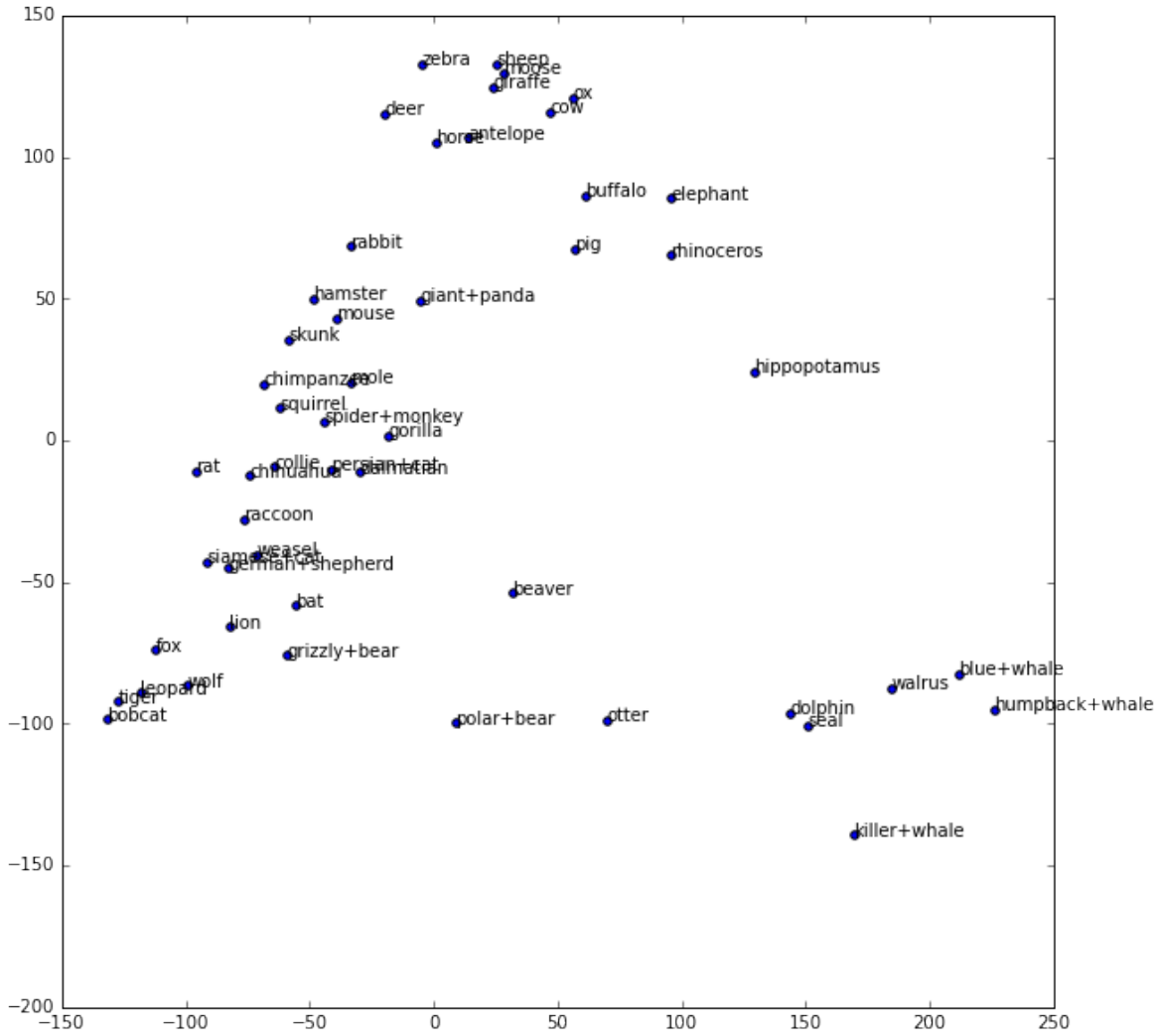


Figure 3: 2D visualization of animals

# 5 Projections

## 5.1 a

The dimensions are $p \times 2$, $2 \times p$, $p \times p$ and $p \times p$ respectively.

## 5.2 b

I think (1) and (3) are the same while (2) and (4) are identical. (1) and (3) are the coordinates of given point x in the new space spanned by $u_1$ and $u_2$. (2) and (4) are the vector representation of x using U.