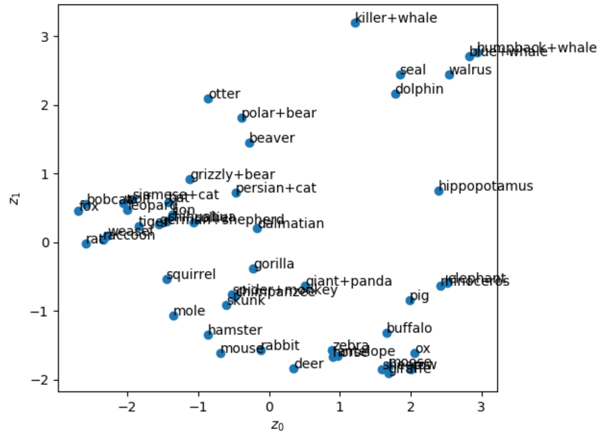
CPSC 340 Assignment 6 (due Friday November 30th at 11:55pm)

1 Data Visualization

1.1 PCA for visualization

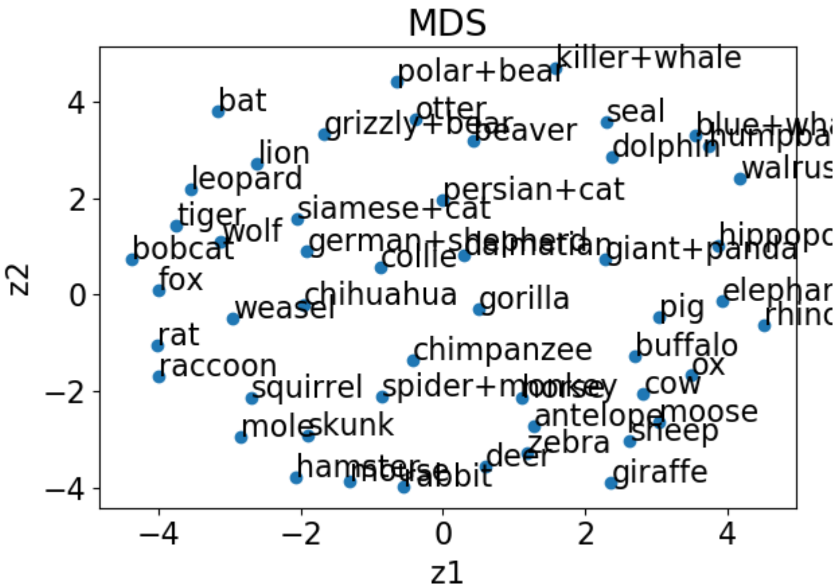
The points are scattered all over the plot on z_0 and z_1 axis.



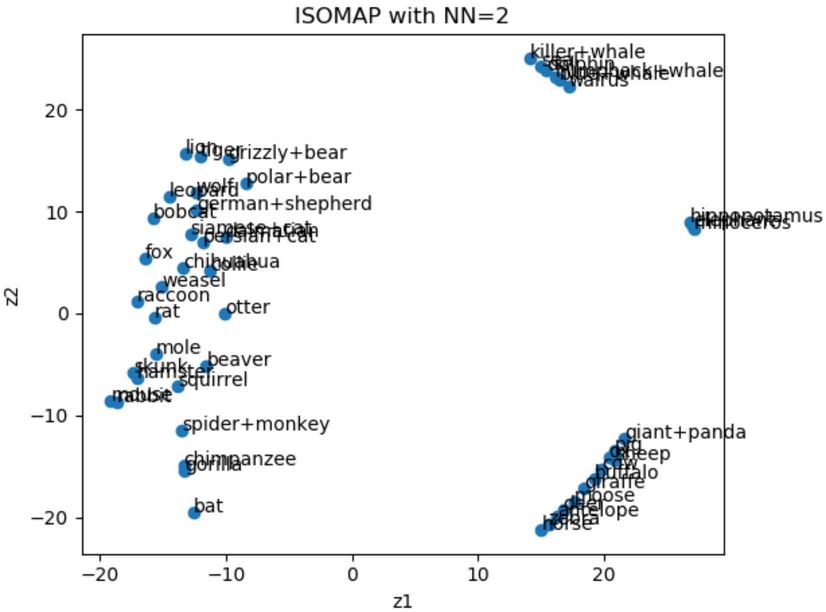
1.2 Data Compression

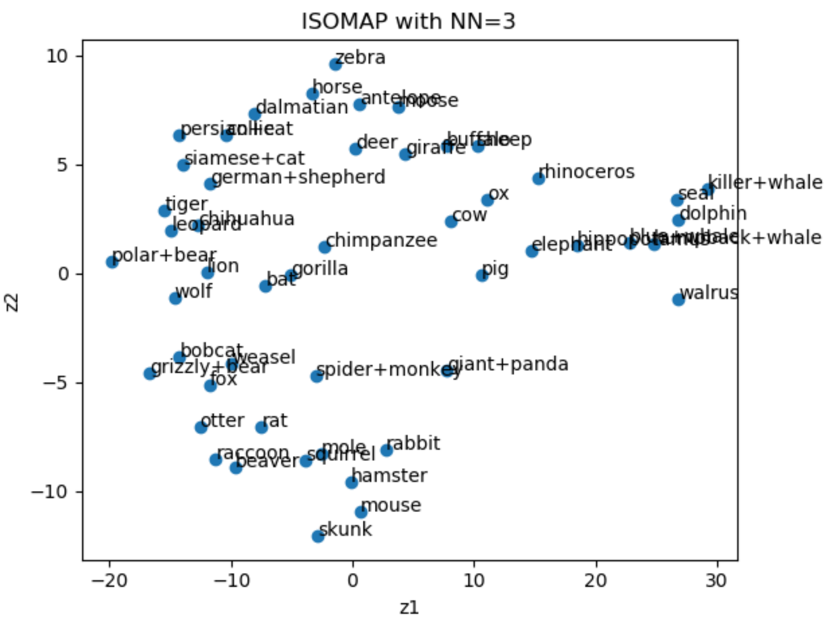
1. 30.19% of the variance is explained
2. We need at least $k \geq 5$.

1.3 Multi-Dimensional Scaling

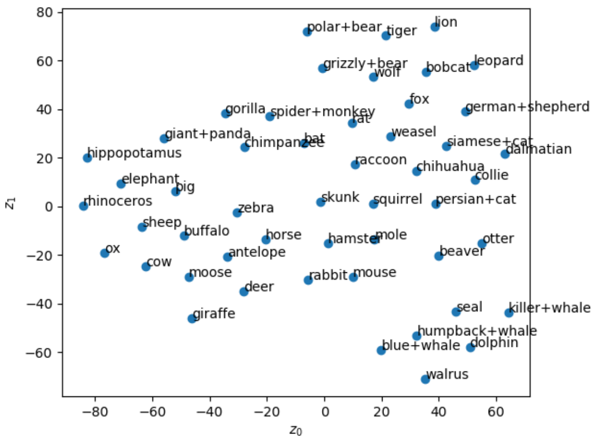


1.4 ISOMAP


ISOMAP with NN=2
z2
z1
20
10
0
−10
−20
−20
−10
0
10
20
killer+whale
walrus
grizzly+bear
polar+bear
leopard
german+shepherd
bobcat
hippopotamus
fox
weasel
raccoon
rat
otter
mole
beaver
squirrel
spider+monkey
gorilla
bat
giant+panda
sheep
buffalo
giraffe
moose
antelope
horse


ISOMAP with NN=3
z2
z1
10
5
0
−5
−10
−20
−10
0
10
20
30
zebra
horse
moose
dalmatian
deer
siamese+cat
german+shepherd
rhinoceros
killer+whale
seal
ox
dolphin
tiger
cow
chimpanzee
elephant
polar+bear
lion
gorilla
pig
wolf
walrus
bobcat
weasel
grizzly+bear
spider+monkey
giant+panda
fox
otter
rat
rabbit
raccoon
beaver
squirrel
hamster
mouse
skunk

1.5 t-SNE



All the plots have relatively scattered points on the graph so it is difficult to differentiate them. The best method would be ISOMAP with $k = 2$ because it does a relatively good job of grouping together the animals.

1.6 Sensitivity to Initialization

MDS is sensitive to initializations and so is t-SNE because t-SNE is a special case of MDS. In the graphs, we see that this is true for t-SNE, which matches up with what we said in lectures. Our MDS is not sensitive to initializations because our implementation initializes MDS the same way every time.

2 Neural Networks

2.1 Neural Networks by Hand

$$z_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, h(z_i) = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{1+e^{-1}} \end{bmatrix}, y_i = \left[\frac{3}{2} + \frac{1}{1+e^{-1}}\right].$$

2.2 SGD for a Neural Network: implementation

Training error = 0.08624 Test error = 0.0839

2.3 SGD for a Neural Network: discussion

I think that the SGD implementation is definetely better as its run time is a tenth of the run time of GD implementation. Although GD is about 3% better in terms of accuracy, it is just too slow for large datasets like this one.

2.4 Hyperparameter Tuning

hidden layers = 100, alpha = 0.001
Training error = 0.00546 Test error = 0.0276

hidden layers = 100, alpha = 0.01
Training error = 0.00144 Test error = 0.0216

hidden layers = 100, alpha = 0.1
Training error = 0.00684 Test error = 0.0212

hidden layers = 1, alpha = 0.0001
Training error = 0.59898 Test error = 0.605

hidden layers = 200, alpha = 0.0001
Training error = 0.0074 Test error = 0.0211

hidden layers = 300, alpha = 0.0001
Training error = 0.00656 Test error = 0.0198

hidden layers = 400, alpha = 0.0001
Training error = 0.00636 Test error = 0.0204

3 Very-Short Answer Questions

1. Yes
2. We can enforce regions that are convex and we may want elements of w to be non-negative.
3. ISOMAP is mainly used for unsupervised learning and it is non-parametric as it first finds the nearest neighbours similar to that of KNN.
4. Content-based filtering is better for new users because it can extract features of users to build a model and use that to predict for new users. Collaborative filtering is bad for new users because they have not rated anything so it will be hard to predict.
5. Collaborative filtering uses L2 regularization and uses stochastic gradient descent.
6. Neural Networks are used for supervised learning to learn features z_i and it is parametric.
7. Regularization helps with reducing overfitting as we add more layers.
8. IDK
9. $10 \times 100 + 100 \times 3 + 100 + 3 = 1403$
10. The neural network will be convex as long as there are no hidden layers. For example, hidden_layer_sizes = 0, activation = relu, solver = adam, alpha = 0.0001, batch_size = 200, learning_rate = constant, number_of_epochs = 10, weight_initialization = random
11. The vanishing gradient problem is that the parameters may underflow or overflow during the calculation of the gradient.
12. Convolutional neural networks has weight matrices that have a particular pattern of sparsity which reduces the number of parameters.