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

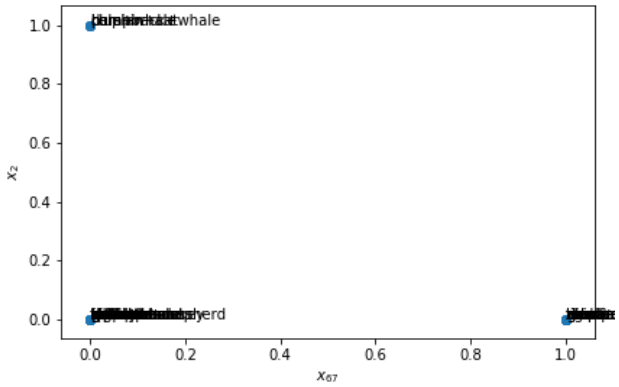
Instructions

Rubric: {mechanics:5}

The above points are allocated for following the general homework instructions.

1 Data Visualization

If you run `python main.py -q 1`, it will load the animals dataset and create a scatterplot based on two randomly selected features. We label some points, but because of the binary features the scatterplot shows us almost nothing about the data. One such scatterplot looks like this:

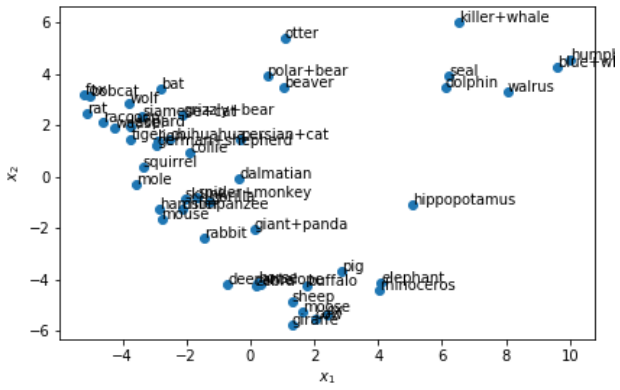


1.1 PCA for visualization

Rubric: {reasoning:2}

Use scikit-learn's PCA to reduce this 85-dimensional dataset down to 2 dimensions, and plot the result. Briefly comment on the results (just say anything that makes sense and indicates that you actually looked at the plot).

Answer: The result is shown as below. From the figure, we can see that the examples are distributed in the whole area of the figure, but not only centered in four points. Thus can give us more information about the animals.



1.2 Data Compression

Rubric: {reasoning:2}

1. How much of the variance is explained by our 2-dimensional representation from the previous question?
2. How many PCs are required to explain 50% of the variance in the data?

Answer:

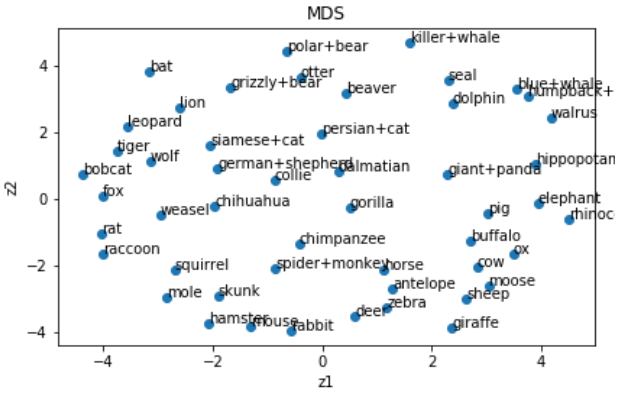
1. The variance is 30%
2. We need to have 5 PCs to explain 50% of the variance

1.3 Multi-Dimensional Scaling

If you run `python main.py -q 1.3`, the code will load the animals dataset and then apply gradient descent to minimize the following multi-dimensional scaling (MDS) objective (starting from the PCA solution):

$$f(Z) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(\|z_i - z_j\| - \|x_i - x_j\|)^2. \tag{1}$$

The result of applying MDS is shown below.



Although this visualization isn't perfect (with "gorilla" being placed close to the dogs and "otter" being placed close to two types of bears), this visualization does organize the animals in a mostly-logical way. Compare the MDS objective for the MDS solution vs. the PCA solution; is it indeed lower for the MDS solution?

Answer: Yes, is indeed lower for the MDS solution. For MDS, the value is about 3553, for PCA, the value 12343.

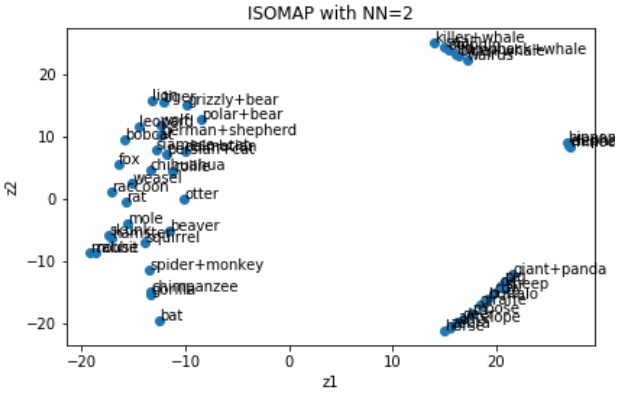
1.4 ISOMAP

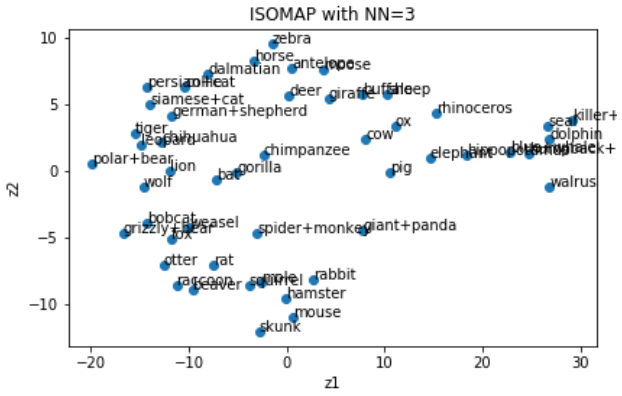
Rubric: {code:10}

Euclidean distances between very different animals are unlikely to be particularly meaningful. However, since related animals tend to share similar traits we might expect the animals to live on a low-dimensional manifold. This suggests that ISOMAP may give a better visualization. Fill in the class *ISOMAP* so that it computes the approximate geodesic distance (shortest path through a graph where the edges are only between nodes that are k-nearest neighbours) between each pair of points, and then fits a standard MDS model (1) using gradient descent. Plot the results using 2 and using 3-nearest neighbours.

Note: when we say 2 nearest neighbours, we mean the two closest neigbours excluding the point itself. This is the opposite convention from what we used in KNN at the start of the course.

The function *utils.dijskstra* can be used to compute the shortest (weighted) distance between two points in a weighted graph. This function requires an $n \times n$ matrix giving the weights on each edge (use 0 as the weight for absent edges). Note that ISOMAP uses an undirected graph, while the k-nearest neighbour graph might be asymmetric. One of the usual heuristics to turn this into a undirected graph is to include an edge i to j if i is a KNN of j or if j is a KNN of i. (Another possibility is to include an edge only if i and j are mutually KNNs.) Answer: please refer to the code for details. The figures are shown as follows:





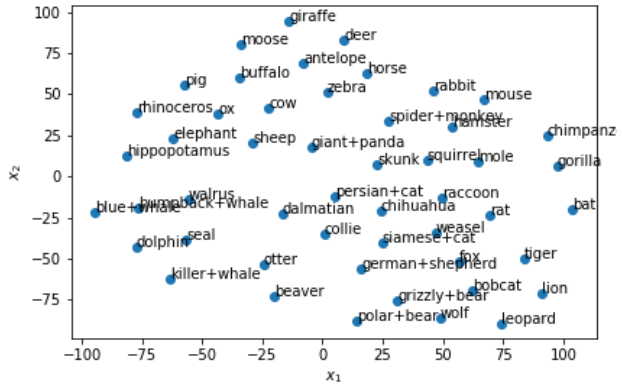
1.5 t-SNE

Rubric: {reasoning:1}

Try running scikit-learn's t-SNE on this dataset as well. Submit the plot from running t-SNE. Then, briefly comment on PCA vs. MDS vs. ISOMAP vs. t-SNE for dimensionality reduction on this particular data set. In your opinion, which method did the best job and why?

Note: There is no single correct answer here! Also, please do not write more than 3 sentences.

Answer: Refer to the figures, we see that ISOMAP with two neighbors has the best performance for dimensionality reduction.



1.6 Sensitivity to Initialization

Rubric: {reasoning:2}

For each of the four methods (PCA, MDS, ISOMAP, t-SNE) tried above, which ones give different results when you re-run the code? Does this match up with what we discussed in lectures, about which methods are sensitive to initialization and which ones aren't? Briefly discuss.

If we run codes several times, we see that t-SNE is sensitive to initialization. MDS is nonconvex thus sensitive to the initialization. This behavior is not reflected from our results if we run several times, it is because that the initialization is fixed.

2 Neural Networks

NOTE: before starting this question you need to download the MNIST dataset from `http://deeplearning.net/data/mnist/mnist.pkl.gz` and place it in your *data* directory.

2.1 Neural Networks by Hand

Rubric: {reasoning:5}

Suppose that we train a neural network with sigmoid activations and one hidden layer and obtain the following parameters (assume that we don't use any bias variables):

$$W = \begin{bmatrix} -2 & 2 & -1 \\ 1 & -2 & 0 \end{bmatrix}, v = \begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

Assuming that we are doing regression, for a training example with features $x_i^T = \begin{bmatrix} -3 & -2 & 2 \end{bmatrix}$ what are the values in this network of z_i, $h(z_i)$, and $\hat{y}_i$?

$$z_i = Wx_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2}$$

Then we have

$$h(z_i) = \begin{bmatrix} \frac{1}{2} \\ \frac{e}{1+e} \end{bmatrix} \tag{3}$$

Then we have

$$\hat{y}_i = v^T h(z_i) = \frac{3}{2} + \frac{e}{1+e} \tag{4}$$

2.2 SGD for a Neural Network: implementation

Rubric: {code:5}

If you run `python main.py -q 2` it will train a one-hidden-layer neural network on the MNIST handwritten digits data set using the `findMin` gradient descent code from your previous assignments. After running for the default number of gradient descent iterations (100), it tends to get a training and test error of around 5% (depending on the random initialization). Modify the code to instead use stochastic gradient descent. Use a minibatch size of 500 (which is 1% of the training data) and a constant learning rate of $\alpha = 0.001$. Report your train/test errors after 10 epochs on the MNIST data.

2.3 SGD for a Neural Network: discussion

Rubric: {reasoning:1}

Compare the stochastic gradient implementation with the gradient descent implementation for this neural network. Which do you think is better? (There is no single correct answer here!)

2.4 Hyperparameter Tuning

Rubric: {reasoning:2}

If you run `python main.py -q 2.4` it will train a neural network on the MNIST data using scikit-learn's neural network implementation (which, incidentally, was written by a former CPSC 340 TA). Using the default hyperparameters, the model achieves a training error of zero (or very tiny), and a test error of around 2%. Try to improve the performance of the neural network by tuning the hyperparemeters. Hand in a list changes you tried. Write a couple sentences explaining why you think your changes improved (or didn't improve) the performance. When appropriate, refer to concepts from the course like overfitting or optimization.

For a list of hyperparameters and their definitions, see the scikit-learn `MLPClassifier` documentation: `http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html`. Note: "MLP" stands for Multi-Layer Perceptron, which is another name for artificial neural network.

3 Very-Short Answer Questions

Rubric: {reasoning:12}

1. Is non-negative least squares convex?
2. Name two reasons you might want sparse solutions with a latent-factor model.
3. Is ISOMAP mainly used for supervised or unsupervised learning? Is it parametric or non-parametric?
4. Which is better for recommending moves to a new user, collaborative filtering or content-based filtering? Briefly justify your answer.
5. Collaborative filtering and PCA both minimizing the squared error when approximating each x_{ij} by $\langle w^j, z_i \rangle$; what are two differences between them?
6. Are neural networks mainly used for supervised or unsupervised learning? Are they parametric or nonparametric?
7. Why might regularization become more important as we add layers to a neural network?
8. With stochastic gradient descent, the loss might go up or down each time the parameters are updated. However, we don't actually know which of these cases occurred. Explain why it doesn't make sense to check whether the loss went up/down after each update.
9. Consider using a fully-connected neural network for 3-class classification on a problem with $d = 10$. If the network has one hidden layer of size $k = 100$, how many parameters (including biases), does the network have?
10. The loss for a neural network is typically non-convex. Give one set of hyperparameters for which the loss is actually convex.
11. What is the "vanishing gradient" problem with neural networks based on sigmoid non-linearities?
12. Convolutional networks seem like a pain... why not just use regular ("fully connected") neural networks for image classification?

Answer:

1. Yes
2. Sparsity improves efficiency and sparsity leads to cheaper prediction and often to more interpretability.

3. unsupervised learning and non-parametric
4. Content-based filtering. Since we do not know the new user's rating but we may know the feature of the user
5. In PCA, $x_{i,j}$ is from real data feature and we pick z and w to approximate the original data x. By using PCA we can achieve the dimension decrease. In collaborative filtering, z and w represent latent features from users and movies separately, and we use z and w to find the label for the user or the movie. The dimension increase in collaborative filtering.
6. Supervised learning and parametric
7. The approximation error will increase when we add more layers due to overfitting. Therefore, the regularization is needed.
8. Because we have a huge set of steps to take, and it does not make sense to check each update. What's more, we are taking many tiny steps and those steps can cause the loss to go up and down. As long as the final error converge, it is fine.
9. parameter = (10+1)*100+(100+1)*3 = 1403
10. We could user Relu as the nonlinear function
11. The gradient will be almost zero for points away from origin. This problem is more obvious when we have sigmoid.
12. For convolutional networks, each neuron only connects to a few nearby neurons while for fully connected layer, each neuron connects to every neuron in the previous layer. Considering that most features in the images are local, using convolutional is more efficient.