Mark Torres
CPSC 453: Machine Learning in Biology
Professor Krishnaswamy
31 October 2019

Project Proposal

Project Idea: Exploring dimensionality reduction and clustering algorithms in the context of image classification.

Brief Idea: Broadly, I'd like to use image classification to learn more about the techniques used in class. In particular, I'd like to perform some of the tasks reviewed (e.g., dimensionality reduction, classification, clustering, etc.) by using image classification as a medium.

Inspiration for project: I don't have much experience working with image data, and I would like to develop those skills. Because image data is high-dimensional, it seems like it would be an appropriate use of the methods reviewed in class.

Overview of computational methods that I plan to implement:

Dimensionality reduction
For dimensionality reduction for images, PCA is a common method used. I would like to perform PCA and compare how it performs to diffusion maps in dimensionality reduction. I would like to then visualize the results of each to show the images before and after reduction.

Image segmentation
For image segmentation (i.e., extracting the different "objects" in an image), I would like to use t-SNE, Louvain, and spectral clustering. I would then compare the original and segmented images to evaluate performance, with both metrics and visuals.

Feature extraction
I would like to use a neural net to extract features from images. This is an addition that I would like to include largely out of self-interest, although I know that deep learning methods are reviewed in the class.

Image Dataset: I am exploring several possibilities for the dataset that I could evaluate. Here are some that I am considering:

- COCO: A well-documented dataset for object classification and segmentation:
- ImageNet: A large image database for common objects
- CIFAR-10: A database of 60,000 images, in 10 classes (good for clustering)

Given the size of these databases, I would use a subset of the images available. I am leaning towards using either the COCO or CIFAR-10 databases.