

Home Assignment #1

Due date: Sunday, May 13th

1. Dimensionality reduction

- a. PCA is used to reduce the dimensionality of a data set. Explain what is the semantic of the first axis of the data after PCA was applied. (5)
- b. Assume that you have a dataset of just with 6 data points in R^3 . And assume that you want to reduce the dimension to R^1 . Construct an example (i.e. pick points) which shows that PCA is would not help you. (5)

2. Density estimation

- a. In the lecture we saw the EM algorithm. EM algorithm is used to estimate the latent parameters of probability distribution given the data. EM is based on a maximum likelihood solution to the parameter estimation problem for a single Gaussian.

In the lecture we saw how to develop the MLE estimation for a single Gaussian in the case of fixed variance. Extend the derivation you saw in the lecture to the case where the variance is also a parameter that needs to be estimated. (13)

3. K-means algorithm:

In the lecture we saw the k-means algorithm for clustering. K-means assumes that the data is given as points in R^d . In this question we will describe the case where instead the data arrives a set of similarities, i.e. a table where for each pair of data points we are given point similarity.

- a. Suggest an approach that would allow using k-means even with this kind of data. Explain any assumptions on the data that you had to make for your approach to work. (10)
- b. A classical algorithm which is designed to handle pairwise data is the **k-medoid**. This algorithm seeks to find a set of cluster representatives (medoid) in the dataset and assigned other items to them. The algorithm randomly picks a k-set of medoids from the data and assigns points to each medoid based on their L1 distances to that medoid. Then, it iteratively tries to improve the assignment by swapping assigned medoid points with non-medoid points so that it tries to minimize the energy of the entire system (which is measured by the sum of distances between medoid points and their assigned data points)
Based on this sketch write down the algorithm and explains why it converges.

Write your answer as pseudo-code and clearly highlight each step of the algorithm. You can use the k-means implantation that you saw in class to guide your answer. (10)

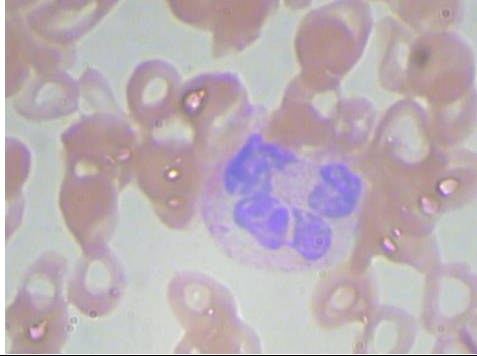


- c. Explain why the **k-medoid** algorithm is more robust to noise then the **k-means** algorithm (7)


4. Clustering:

In this question you are given a dataset of blood images i.e. white and red colored blood cells. While red blood cells are relatively similar, white blood cells come with different shapes and sizes.

Below you can see sample images showing the different types of white blood cells which correspond to different clusters.

Your task is to cluster the data set so that similar white blood cells are clustered together.

	NEUTROPHIL	train_00000.jpg
	LYMPHOCYTE	train_00021.jpg
	MONOCYTE	train_00015.jpg

	EOSINOPHIL	train_00009.jpg
---	------------	-----------------

You are given a python file which has two skeleton methods, a Train method and a Test method. The Train method receives a list of image file names that should be used for training. This method constructs a segmentation model and return this model. The Test function receives a model that was previously trained by the Train method and a list of image file names. It uses the model to produce a list of class ids and return this list, i.e. all images that share the same cluster should have the same id.

When you submit your code for this part of the exercise make sure to send a single file with the following naming convention <ID>_ex1.py. Your code is allowed to use everything that is in python's standard libraries as well as everything that comes with numpy, scipy, sklearn or opencv. You are allowed to add as many helper functions as you want but you need to retain the interface that is defined in the skeleton code.

In addition, you should submit a python notebook where you invoke your algorithm and visualize the results. You can simply copy the code from <ID>_ex1.py. to the notebook and evaluate it there. The notebook should be named <ID>_ex1.ipynb.

When you submit, wrap these two files into a single zip file named - <ID>_ex1.zip

The data for this exercise resides in a file named ex1-data.zip.

When we check this question, we will first execute the Train method and then test the results in the test method with data you are not given a-priori.

- Describe the feature set that you are using, justify the different features you decided to use. (15)
- Describe the different clustering algorithms you tested and explain the reasons for the selecting the algorithm that you submitted. (15)
- Submit the code as was described in the preface of this question. (20)