Office hours

# #7

## Piazza Questions

* *The 5th step: "Apply the clustering algorithms to the same dataset to which you just applied the dimensionality reduction algorithms (you've probably already done this), treating the clusters as if they were new features. In other words, treat the clustering algorithms as if they were dimensionality reduction algorithms. Again, rerun your neural network learner on the newly projected data."*

*Want to confirm whether this step means:  
1) clustering -> neural network  
 OR  
2) dimension reduction -> clustering -> neural network*

* *1. Can you give some advise on how to describe the clusters that result from the algorithms?  
  2. What can we use to define "good" and "bad"clusters?  
  3. In part 5 "apply the clustering algorithms to the same dataset to which you just applied the dimensionality reduction algorithms, treating the clusters as if they were new features. In other words, treat the clustering algorithms as if they were dimensionality reduction algorithms. "  
  Does this mean, my dataset has 10 attributes + 1 label columns. Now I add 1 attribute, which is the cluster index as a new column (so now I have 11 attributes + 1 label)?   
  I am confused by this part "treat the clustering algorithms as if they were dimensionality reduction algorithms".  
  4. What are some "interesting" findings we should look out for?*

## End Questions



# #8

## Piazza Questions

* *when there are categorical variables (0 or 1 in my case), how should we treat them in clustering and dimension reduction? Shall I just keep 0 & 1 for categorical and other continuous variables are normalized?  
  I see somewhere mentioned that PCA is not suitable for data with categorical variables and something like CATPCA is recommended? Should we go ahead using PCA for the dataset with categorical variables as instructed?*
  + Should be able to see in the output. “This will certainly be interesting from, say, analyzing the distribution of the eigenvectors in PCA.
* *From assignment:clarification about words in double quotes:*

*"Why did you get the clusters you did? Do they make "sense"? If you used data*

*that already had labels (for example data from a classification problem from assignment #1) did the clusters line up with the labels? Do they otherwise line up naturally? Why or why not?"   
"Sense" meaning the following questions.  
Basically, students are supposed to determine whether or not the clusters line up with the labels.  
This is "sense" in the current context.  
Q2:Is this understanding correct?  
"Do the projection axes for ICA seem to capture anything "meaningful"?"*

*Q2.Please define "meaningful" in this context.*

* *For assignment three, when running the clustering algorithms, can we specify K just because we know the number of classes in our data from assignment 1? I hope that is a fair enough justification. i.e my data has 10 classes so i will try the clustering with k=10.*
* *A3:Clustering :Hyperparameter tuning   
  The value of k i.e number of clusters for kmeans is a hyperparameter.  
  For A1 ,which was in the domain of supervised learning , hyperparameter tuning was done using model complexity analysis.e.g validation\_curve in sklearn.  
  However,A3 is related to unsupervised learning.  
  Consequently, the supervised learning metric cannot be used here.  
  Questions:  
  Q1.What are possible ways to perform hyperparameter tuning in context of unsupervised learning?  
  Q2.Do learning curve and model complexity concepts transfer over from supervised learning to unsupervised learning ?*
* *For assignment 3, has anyone come up with a good way to explore the clusters?  
  It looks like Silhouette and BIC are the right ways to pick the number of clusters for K-Means and EM, respectively.*[*https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6*](https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6)
  + <https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b>
* *A3:how to think in terms of picking the right dataset for A3?does a binary classification dataset make sense ?*
  + Having a non-trivial amount of features, I think, that you can apply dim. red. to is also important.
* *"For PCA, what is the distribution of eigenvalues" Is it refer to histogram of eigenvalues or something like cumulative eigenvalue (explained variance) against number of principle component?*
* *Question @567 How to get ground truth for clustering performance metrics (homogeneity score and others)? I've got optimal number of clusters K=7 for my data set and I see that it's correct, based on graphical representation of the clusters, but the target is binomial, so I cannot use it as a ground truth for the clustering evaluation metrics with this data set, what can I use instead a s a ground truth in this case?*
  + "Ground truth" is the definition of the first param for homogeneity\_score() API
  + Just to add a little bit from the @567:   
    >Non-rigid bodies such as humans can however produce multiple clusters per target. Therefore, the number of clusters does not always remain equal to the number of targets. Hence, calculating the number of targets on the basis of the number of clusters can be inaccurate...
* *How to choose number of components in ICA/Randomized Projection?   
  Has anyone come up with a good way of choosing the number of components in ICA and Randomized Projections? I am using sklearn. The PCA module in sklearn seems to let you give it a float for the number of components:  
  ```If 0 < n\_components < 1 and svd\_solver == 'full', select the number of components such that the amount of variance that needs to be explained is greater than the percentage specified by n\_components.```*[*https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html*](https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html) *ICA doesn't seem to have anything similar, and I don't see an easy way in their documentation to get information about how independent each feature is. Maybe we could try to get this from the mixing matrix, but I haven't been able to figure that out yet. SKLearn docs:* [*https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html*](https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html) *For ICA, does it make sense to do the following:  
   Run fit\_transform on the x\_data with FastICA having n\_components values ranging from 1 to the number of features  
   Calculate the absolute value of kurtosis of each feature  
   Take the mean of all these values, and use that as a score for the FastICA algorithm with that number of components  
   Pick the number of components that maximizes this score?*
  + There is no hard and fast rule to doing this, but I think you have a good start here on one possible way. I think what you're getting at is using the highest average kurtosis to pick the number of components.   
    You should think about what this practice means, as well, in terms of non-Gaussianity and maximizing independence. In addition to that, you may supplement other metrics that deal with Gaussianity or look directly at the probability density plots to inform your decision on the number of components.  
    This paper seems to have helped students a lot with getting started.  
    <https://www.cs.helsinki.fi/u/ahyvarin/papers/NN00new.pdf>
* *For A1,A2 as well as A3 neural networks seem to be a common component. My understanding for them to be common is because they are relatively harder to train given the larger number of parameters to tune :learning rate, momentum, hidden layer size etc. Is this understanding right? if not, please clarify the reasoning if any. Mostly curious hence this Q.*
  + For A2, we use the randomized algorithms to optimize the weights in the network, so it's, most likely, just following that pattern of carrying that original baseline forward to this assignment. NNs can also learn highly non-linear functions, many of which you all choose, so they perform well enough generally to provide a decent baseline.
* *In the instruction of Assignment 3, it asks us to include "a description of the kind of clusters that you got". What does "the kind of clusters" mean? What kinds of clusters are there?*
  + After you've run clustering and chosen K, you should work on providing metrics or visuals that describe the quality of the clusters. So we're really asking how dense those clusters are, separated they are, whether they correspond to your original labels or something else, et cetera. There are many ways that you can go about this, so we just ask you to justify whatever metric you use
* *Hi, I read few piazza posts too. But still unclear. What is clustering supposed to do? I was assuming it would logically group data into clusters according to its attributes. my dataset has binary classification.. 120+ attributes.. clusters 60+ are showing small WCSS, which is good. But does this mean that it is good to have 60+ clusters for this dataset? am not sure if am doing right. how many clusters r we supposed to expect? I read in other posts that # of clusters <> No of classifications of labels. So what is the right/optimum value (I understand the elbow rule, but trying to understand practically what # of clusters mean). Sorry, if am repeating similar question, but other answers were not clear for me.*
  + So, the number of clusters you get may not necessarily correspond to your original labels. But that's interesting for your analysis... you should move forward by trying to explain that. I would look at the cluster centers output from K-means, for example, and try to relate that back to the features to start if you're not looking purely at the quality of the clusters.   
    Depending on the data set, your unsupervised clustering may have found some explainable (via your features) signal in your data that is different from the original labels.

## End Questions



# #9

## Piazza Questions

## End Questions