**Lesson 04 Discussion**

As you have seen in class and in your lab, unsupervised learning models are useful in finding relationships in unlabeled data. Since much, and perhaps most, of the available data in the real world is unlabeled, unsupervised learning methods are often quite useful.

Think about a machine learning problem (real, or hypothetical) where you might apply unsupervised learning. Discuss the following points:

1. Why is unsupervised learning useful for this problem?
2. Which unsupervised learning methods would you plan to try, and why?
3. How will you evaluate and compare the results of your models?

  To receive credit, post you thought and then comment on at least one other post.

As I mentioned before in a previous discussion, one of the ways I have applied unsupervised learning was to explore thermodynamic data on commercial engines to find clusters/patterns as well as a way to determine feature importance on a prediction model.

<https://www.ibm.com/cloud/learn/unsupervised-learning>

<https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a>

One way in which I used unsupervised learning before was to look for clusters inside a particular type of commercial engine to see if we could customize the rework/overhaul done to the engines based on this clustering analysis. The main reason for this was to optimize our resources given than previously we were assigning the same type of rework/overhaul to all engines regardless of their actual need. By using unsupervised learning, we were hoping to determine specific customize levels of rework/overhaul and by doing this optimizing the resources of each test cell/overhaul shop.

At that time, we decided to use k-nearest neighbor for this analysis, as we were looking for an easy model to implement and give us a clear distinction between groups (if any). One of the critical points during this analysis was choosing the right “k” value (number of clusters) as we were exploring the data. For this cross-validation was critical.

For this project the way it was decided to evaluate the model was by taking a period of time when the new engines could go to the classification model using these newfound clusters and once the rework was actually performed on the engine, we checked against the unsupervised prediction to see how far or close it was to the real event. As mentioned before this analysis/project was done to optimize the material/people that each test cell/overhaul shop actually needed based on the type engines there were getting. By doing this (plus a margin of safety) we were able to cut on some of the resources we were wasting by assuming the worst case scenario for all engines.

there was correlation between this grouping and the type of rework/overhaul we would need to do to the engines to get them to the desire performance level.

obtain/analyze different clusters on the same kind of commercial engine type to see if there were any patterns that we could potentially use to customize the rework/overhaul that we were doing to the engines. In

*While working on the thermodynamics’ side for commercial turbofan engines very quickly we realized that if we wanted to have any prediction model in production looking at “on-wing” and/or “test-cell” data providing real time predictions we needed to somehow reduce the number of features entering the model while maintaining certain degree of control and accuracy. As you can imagine for turbofan engines we have many sensors while flying, plus extra measurements from extra sensors while testing the engine on the ground (“shop-test cell”), which in turn might lead to thousands of inputs for every single data point. During this time many different methodologies were implemented to try to solve this problem of dimensionality reduction and one of those techniques implemented in some of the models (especially for those with a high degree of complexity) was PCA to find a lesser set of features that could provide a comparable (or even better) performance, while improving the time response. Using this methodology we were able to find great insights on the contribution of the kind of sensors we were using, but ultimately for most of our models this technique was just the first step for feature selection and not the final answer to our high dimensionality issue. This mainly because our customers needed not only the prediction but a model that was easier to interpret.*