**0 – Introduction to Unsupervised Learning**

Statistical or Machine Learning in general refers to a vast set of tools for understanding data. These tools can generally be classified as either supervised or unsupervised methods. In supervised statistical learning (DSCI 425) our interest is to develop statistical models to predict a response () using set of predictors (. In unsupervised learning we generally do not have a response () to predict but rather our goal is to discover interesting things based on a set of measurements (). However, often times data does not even fit into this restrictive format. For example, in a collection of social media posts (e.g. Twitter), user reviews (e.g. movie reviews on IMDB), or images, it is not clear how these “data” could be put into () format.

Here is a simple example taken from dataaspirant.com.



* Suppose you have a basket filled with different kinds of fruits: apples, bananas, grapes, and cherries.
* Your task is to arrange them into groups.
* The fruits in the basket are as follows:

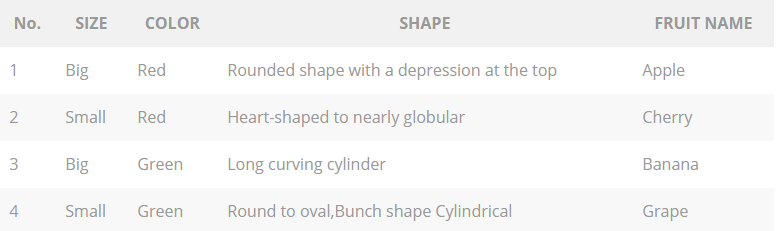
Apples Bananas (unripe) Grapes Cherries

**Supervised Learning**

As we already know a lot about these fruits we can define measurements that will allow us to classify the four types of fruits. Our training data would consist of the relevant measurements/characteristics () we could use the classify the type of fruit (response ). The table below gives a model that can be used classify a fruit according to its type. Given a new piece of fruit of one of these types we could use the rules below to classify the fruit type accurately. Of course some apples are green, ripe bananas are yellow, and not all grapes are green, so we have to update our rule to incorporate these new situations as more data becomes available.

Obs.



The key is we have a set of measurements (size, color, shape) and a target response (fruit type). Our goal in a supervised learning problem is develop a model/rule to use the given to accurate predict the response .

**Unsupervised Learning**

We now consider the same scenario, but this time we assume we don’t know anything about the fruits (i.e. this honestly the first time you have ever seen them). So how might we go about arranging the fruits? For example, we might decide to group them on the basis of color.

* Red Color Group: apples and cherries
* Green Color Group: bananas (unripe) and grapes

If we take another characteristic into consideration such as size then our groups might become:

* Red Color and Big Size Group: apples
* Red Color and Small Size Group: cherries
* Green Color and Big Size Group: bananas
* Green Color and Small Size Group: grapes

The rule ends of the same as the supervised learning problem, but this time we did not have a response variable we were trying to predict accurately, i.e. fruit type. Here we did not even know apples, cherries, bananas, and grapes even existed. We simply took the measurements given and organized the observations (fruits in the basket) according to these measurements. In an unsupervised problem we don’t have a response () we are trying to predict accurately – rather we use only to uncover structure in our data.

As mentioned above our data may be in a format that does not lend itself to the standard *rows = observations* and *columns = variables* format. For example, each observation might consist of a tweet on Twitter ® or the text found in restaurant review on Trip Advisor ®. Thus the data we are working with may require a lot of pre-processing in order to get it in a usable format for the different methods we will be examining. This may also require developing *metrics* to measure data features of interest, e.g. counting the number of words used in a restaurant review that are indicate a favorable or negative sentiment about a restaurant.

Below are some examples of the types of problems that can be examined using unsupervised learning methods, many of which we will be seeing in this course.

1. Given a large customer database, find groups of customers who have similar traits. Once we have our customers divided in these groups or *segments* we can potentially employ different marketing strategies to increase sales or introduce new products to target specific groups in the customer base. This is called customer or market segmentation, but is really just an example of a broader method called ***cluster analysis***.
2. Given an individual Netflix ® subscribers reviews of movies they have watched, can we recommend other movies that might also like based on these reviews. This was the goal of the *Netflix Prize* you may have heard about. This is typically done by aggregating or averaging reviews of other subscribers with determined to have similar tastes in movies for films/shows the subscriber has not yet seen or rated. This is the general idea behind ***recommender systems*** which is a topic in this course. Amazon ® employs an excellent recommender system when suggesting products its online shoppers.
3. Given a set of reviews, comments or tweets about something (e.g. a restaurant, a new product, or a political candidate etc.) can we judge the general sentiment of the populace? This is the general idea behind a method we will be discussing called a ***sentiment analysis***. As another example, a Kaggle ® problem a few year back asked users to develop methods to accurately score the written portion of the ACT exam. This problem is somewhat supervised in that the response (Y) was the score already assigned to the body of text by ACT, but this process does involve some methodology we will be examining in this course.
4. Large retail stores are often times interested in what products are being purchased together. For example, a grocery store might want to know if a customer purchases milk are they also likely to buy breakfast cereal? Knowledge about which products tend to be purchased together can be helpful when designing store layouts, product placement, and sales promotions. This is sometimes referred to as market basket analysis or more generally ***association rule*** analysis.
5. Given a set of numeric measurement ( where is “large”, what can say about the correlation structure amongst these variables? Here “large” may be situations where or possibly even more! Often times when is this large, it will typically exceed the sample size . Understanding the correlation structure or in general multivariate structure in set a numeric variables is typically done using dimension reduction methods such as ***multidimensional scaling (MDS)****,* ***principal component analysis (PCA)****,* ***factor analysis (FA).*** Which of these methods is employed depends on the goal of the analysis and how the results will be used. For example, one use of principal component analysis is to simplify the predictor space when using developing regression models.
6. Suppose multiple sensors placed on the head are used to measure brain signals for a sample individuals who are asked to perform a series of motor functions like clenching their teeth, tapping their fingers on a table, rolling their eyes, etc. As these sensors are placed are all over the surface of a subject’s head, each sensor is receiving electrical impulses from all over the brain. Despite the fact that each sensor is potentially receiving impulse information throughout the brain, the strongest signal is probably obtained from the portion of the brain nearest the sensor. Can the measured signals from these multiple sensors be used to break them down into independent components that represent the separate distinct signals coming from different regions of the brain? This again is a dimension reduction question, however is best solved using a technique called ***independent component analysis (ICA)***.
7. Given the results of a survey what can be said about the association between demographics of the respondents and their responses a set of survey items? Data of this type is typically a mixture of ordinal and nominal variables, i.e. not meaningfully numeric. Examining the structure of this type of multivariate data is typically done using ***multiple correspondence analysis (MCA)*** another topic in this course. MCA is closely related to PCA above.
8. A more general form of unsupervised learning is ***data visualization***. Data visualization is an important part of **all** types of analyses (supervised and unsupervised). We can however view it as type unsupervised learning as we can use graphical methods to learn or identify important features, structures, and relationships in our data. Visualization will be stressed throughout the topics covered in this course as many of the methods use visualization to aid in the interpretation of the results.
9. ***Social network analysis (SNA)*** is the process of investigating social structures through the use of networks and graph theory. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them, e.g. visualizing connections in social media accounts such as Facebook ®. This topic will be covered as time permits.