**1.1 – 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 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 ().

Here is a simple example taken from dataaspirant.com.



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

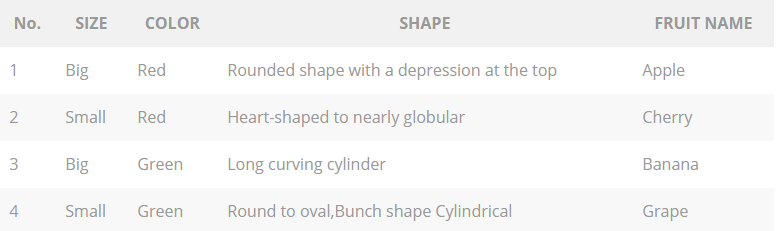
Apples Bananas (unripe) Grapes Cherries

**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 ). Table below gives 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 rule 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.

The set of measurements 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. This is called customer or market segmentation, but is really just an example of broader method called ***cluster analysis***.

1. 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 reviews of other subscribers with similar tastes in movies for movies 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.
2. 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***.
3. 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? What items frequently occur together in a person’s shopping cart and which products are almost always bought in tandem, i.e. if they buy one item they are very likely to also buy the other. This type of knowledge can help with designing store layouts, product placement, and sales. This is sometimes referred to as market basket analysis or more generally ***association rule*** analysis.
4. 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.
5. Given the results of survey of individuals about leisure activities they participate in, what can be said about the relationship between their choices and their sex, age, marital status, and their profession? What activities are most associated with other? What information in the survey best explains inter-individual variability? These are questions that can be answered by using *dimension reduction* techniques such ***as multiple correspondence analysis (MCA)*** which is a special case of ***principal component analysis (PCA)***.
6. Given a set of numeric measurements on individuals, can we construct indices that capture of much of original variation between individuals upon which these measurements were made? For example, if we took a large number of body measurements (arm length, thigh circumference, waist size, etc.) could we develop a single numeric index representing overall body size? This again represents a dimension reduction question which can addressed by using ***principal component analysis (PCA)*** or ***factor analysis (FA)***.
7. 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 each sensor is potentially receiving impulse information throughout the brain, the strongest signal is probably obtain from the portion of the brain nearest the sensor. Can the measured signals from these multiple sensors be used to break them down to independent components that represent separate distinct signals coming from the brain? This again is a dimension reduction question but is best solved using a technique called ***independent component analysis (ICA)***.

**1.2 - Introduction to Supervised Learning (aka Statistical Learning)**

As mentioned previously statistical learning in general refers to a vast set of tools for understanding data. Again in supervised statistical learning our goal is to develop statistical models to predict a response () using set of predictors ( and in unsupervised statistical learning we do not have a response () to predict but rather our goal is to discover interesting things about a set of measurements ().

In ***supervised statistical learning*** we build statistical models for predicting, or estimating, an output (response ) based on one or more inputs (predictors or features ). The response () can either be a numeric or categorical/ordinal variable. If the response is numeric we attempt to build a model that accurately predicts the value of the response (e.g. home price, future sales) using a set predictors. If the response is categorical/ordinal with *k* different levels (e.g. disease type, cancer grade, or whether a credit card transaction is fraudulent or not) then we can develop models using a set predictors that estimate the class an observation belongs to, or in the case of some modeling strategies estimate the probabilities an observation belongs to each of the *k* classes.

Here are some examples of questions where supervised statistical learning could be employed to help answer them.

1. How does the mean selling price of a home relate to the physical features of a home like the # of bedrooms, # of bathrooms, square feet of living space, neighborhood home is located, whether the home has a fireplace or not, etc.? How accurately can we predict what a home will sell for given these characteristics? (Links: [www.zillow.com](http://www.zillow.com) and [www.redfin.com](http://www.redfin.com))
2. Given information about a potential loan applicant (e.g. age, home ownership, annual income, credit score, etc.) can we classify the applicant a good or bad credit risk?
3. Given a large set of genetic expression levels (these are generally numeric) can estimate the disease status of a patient? Can we estimate the probability they will be successfully treated using the current chemotherapy regimen used to treat that type cancer assuming they have it? Can we estimate the probability that someone will develop a certain disease in their lifetime given these genetic expression levels?
4. Given a set of hospital characteristics (e.g. number of orthopedic surgeries per year, # of inpatient beds, # of surgeons, geographic location, etc.) can we estimate the orthopedic sales potential for the hospital?
5. How are the shape and size measurements of breast tumor cells related to the malignancy? Given a set of these measurements can we accurately classify the malignancy of the breast tumor? Can we model the probability that a cell is malignant (or benign) based on these measurements? Can we use these measurements to estimate the probability of tumor recurrence for patients with malignant tumors?
6. Given information about an advertising campaign (e.g. amount spent on radio ads, TV ads, internet ads, length of campaign, etc.) can we estimate the sales impact in dollars?
7. Given characteristics of an e-mail (e.g. word count, capitalization, whether it contains certain keywords, etc.) can we accurately classify it as *spam*?
8. Given information about a credit card transaction (e.g. amount of charge, location of charge, type of purchase, etc.) can we accurately classify it as a legitimate or fraudulent charge?
9. Given a set of information about a chemical process can we accurately predict the chemical yield or some other numeric characteristic of the chemical such as purity or solubility?
10. If Target.com purchases a banner ad on Google for say HD TV’s can we estimate the probability that if someone clicks on it they will actually make a purchase? What is the estimated ROI (return on investment)?
11. If LL Bean mails a catalog to a potential customer can estimate the probability that they will actually make a purchase given customer information (e.g. age, income, previous purchase history, gender, etc.)? If they do make a purchase can estimate amount of their purchase using these same characteristics?
12. Given a set of physical characteristics of a mushroom can we classify it as poisonous or edible?

As you can probably guess the possible applications of supervised statistical learning methods are infinite! Notice that the hypothetical examples above are a mixture of prediction problems where the response is numeric and where the response is categorical/ordinal. Also notice the broad class is application areas (e.g. business, medicine, chemistry, etc.) where these methods can be used.

**Questions and Tasks**

Realizing you have had only a brief introduction to the IPUMS data, can you develop two research questions, one where supervised learning methods and one where unsupervised learning methods could be used?