# Examples of Clustering Techniques:

## K-Means

### Short Comings:

* Undirected, no good estimate of target
* R^2 can be small even when the line appears to be a good fit making the evaluation of the model difficult
* The width of the boundary of the cluster expands in all directions equally potentially combining observations in another group
* The centroid of the k-means clusters aren’t always the best approach when the cluster is between two dense regions (little data points nearby)
* Better for well-defined clusters but hard to do with an automated approach because that information isn’t known before applying the k-means clustering technique

### Some Solutions:

* Transform the data so that the points rotate to form a horizontal line that separates the clusters. Expand the data points vertically to allow for more separation between the clusters
* When transforming the cluster results back to the original data points, the clusters form ellipses that do a better job defining clusters than applying the k-means to the original data set (circles)

## Gaussian Mixture Models (GMM) / Expectation Maximization (EM)

* Think of each cluster as a set of Gaussian probability distributions. The probability distribution gives the probability of a point anywhere in the space belonging to the cluster centered where Gaussian is centered
* Multi-Dimensional Gaussian:
  + In 1-dimension
    - Gaussian distribution represents a bell-shaped probability distribution where the mean is zero and standard deviation is 1
    - Probability is derived by calculating the area under the curve
      * EX: Calculate where the value is negative. Since the normal distribution is centered at 0, the area is .5 or probability = 50%.
  + In 2-dimension
    - Looks like a symmetric hill
    - Probability is derived by calculating the volume under the curve
      * EX: If you need to know how often two variables are negative, calculate the volume where both values are negative. Answer = .25 or 25%
  + The standard deviation for a multi-dimensional distribution can be assigned by estimating the standard deviation that maximizes the likelihood of the points belonging to the cluster.
* Gaussian for K-Means Soft Clustering
  + Result of using Gaussian for K-Means = a model that scores all records to produce probabilities (representing the likelihood) of each record (observation) being in each cluster.
  + The probability of an observation being in a cluster can be calculated by the likelihood of an observation in cluster X divided by the sum of all the likelihoods
  + As a result, each cluster has two assignments:
    - The hard cluster assigned by k-means
    - The soft cluster assigned by a probability of being in each cluster
      * The dominant cluster is the one with the largest probability (sometimes different from the hard cluster)
    - Repeat the process using the dominant cluster as the assignment to calculate the center and standard deviation of the probability distribution. Repeat until a stable set of soft clusters is formed
* Standard deviations for GMM do not have to be the same in all directions which result in ellipses instead of circles
  + This changes the shape of the Gaussian where the symmetric hill (when standard deviations are the same for all variables) now looks more like a mountain where slops can be steeper in one direction than the other
    - In other words, the ellipses can be titled at any angel and be placed anywhere on the plane
* Cross-sections define a particular distribution
* Canonical measure defines the cross-section that is 90% of the distribution
  + i.e. 90% of the volume under the ellipse
* Simplified: GMMs are fitting ellipses to best fit the data points = produce the best probabilities of membership in the clusters
* Expectation Maximization algorithm: Two steps:
  + Expectation Step: Calculates the probability of each point being in each cluster
    - INPUT: Set of cluster centroids and variance parameters for the clusters
    - OUTPUT: Set of probabilities for each data point that represent the data point being in the cluster centered at each centroid.
  + Maximization Step:
    - Recalculate the cluster centroids and distributions based on the assignments of the data points.
    - The center of the cluster = the weighted average of all the points
    - The parameters that describe the distribute are calculated from the data points using statistical formulas.
* Scoring GMM steps
  + Calculate the likelihood that each record belongs to the cluster
  + Normalize the likelihoods to obtain the probabilities of membership in each cluster
    - Divide each one by the sum of all of them so the total sums to one
  + (Optional) assign the cluster membership based on the highest probability.
* WARNINGL GMM clusters are sensitive to outliers
  + Can create a large cluster in the middle and smaller clusters containing outliers.

## Divisive Clustering

* Applies a top down approach to clustering
* Starts by using k-means to split data into two clusters. The largest cluster is split again until some condition is met (like decision tree)
* When comparing to a decision tree, both use a purity formula to separate the data. A decision tree has a target variable, a divisive clustering technique does not.
* Measuring Purity:
  + Measuring Purity on all Numeric Columns:
    - Measure of Tightness: Two resulting nodes are as small as possible.
    - Measure how far apart the nodes are, the further apart, the better
  + Measuring Purity on Categorical Fields:
    - Apply the chi-squared test to all variables, create contingency table for all variables.
* General Approach:
  + Calculate the p value to answer the question: How likely is it that the split on a given column is due to chance?
    - For categorical values, look up p-vale that corresponds to the chi-squared value
    - For numerical, use the standard error of the mean (SEM)
  + The p-value determines if the split is interesting.
* Scoring Divisive Clusters
  + Similar to decision tree where each split eventually will place an observation into a leaf node which will equates to the cluster assignment.

## Agglomerative Clustering

## Self-Organizing Maps - a neural network approach