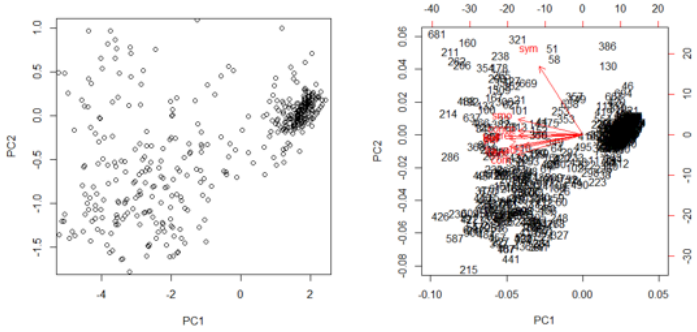
Data Set: For this analysis I have used a dataset sourced from the UCI machine learning repository that contains 699 observations of breast tumors each with 10 attributes. Such attributes describe the tumor using a combination of physical and visual measurements that are given a rank from 1 to 10 by the medical examiner. These variables include symmetry, thickness and smoothness/texture ratings. From the data set we know that observations are divided between benign tumors, 65.5%, and malignant tumors, 34.5%[i]. Since I am interested in testing the accuracy of the clustering algorithms which are explained below, I have removed this result vector from the data set and will use it later to test accuracy.

Since there are 10 independent variables that describe each tumor, I used principal component analysis, which is a dimensionality reduction technique that allows us to plot our observations using only 2 or 3 dimensions. This is useful when plotting on 10 dimensions would be infeasible. PCA produced the following graphs:



Both charts to the right plot observations measured against the first two principle components which represent roughly 80% of the data, determined by a scree plot. The left chart shows the general dispersion of the datapoints with one very distinct cluster near the top right while all other observations are spread on the left. The chart on the right shows the same data dispersion with a superimposed layer of arrow vectors representing each independent variable. This shows that the variable that measures cellular symmetry is the least correlated to other variables in the original data set. This could turn out to be an important factor that determines the diagnosis of a tumor.

Problem: We would like to distinguish subgroups within our dataset that will help us learn more about the difference between benign and malignant tumors. Clustering algorithms are often useful in medical research and cases like ours where we are interested in grouping observations based on their observed characteristics. Often, researchers may not know how many distinct clusters they have in their dataset and clustering algorithms are a helpful statistical tool that can help shed light on this uncertainty.

Techniques: To learn more about our dataset and each observation within it, I have used an unsupervised learning approach referred to as clustering. Clustering seeks to separate observations within a dataset into subgroups and is considered an unsupervised learning approach because instead of trying to "predict" some dependent value we are instead using the relationships within each observation to discover similarities that will help us develop insights into the nature of the problem.

K-Means: The K-Means algorithm is a popular clustering algorithm that works on the idea of creating a predefined number of clusters and iterating until the data points in the data set are assigned to the cluster than minimizes within cluster variation.

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Here, $W(C_k)$ is the within cluster variation where C_k represents the K^{th} cluster. Within cluster variation is defined as the sum of all squared Euclidean distances between each point divided by the total number of observations within the K^{th} cluster. The algorithm works iteratively by first randomly assigning all observations to one of the K clusters. A cluster "centroid" is calculated as the center distance between all points in the cluster. The algorithm continues to iterate and move datapoints between clusters that minimize each cluster's internal variation. The algorithm stops once data points are no longer reassigned between clusters. [ii]

Hierarchical clustering: Hierarchical clustering has the added advantage over K-Means in that a predefined number of clusters is not required for the algorithm to work. Instead, the algorithm builds a tree like structure referred to as a dendrogram. I have used an agglomerative approach to build such a dendrogram which starts by putting each observation into its own cluster then iteratively merges each cluster to its most similar neighboring cluster. This continues until there is only one cluster containing all data points in the set.[iii]

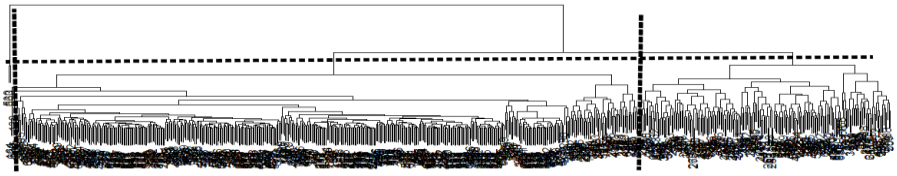
Gaussian mixture models: With a Gaussian mixture model, clusters are modelled as Gaussian distributions. The Expectation-Maximization algorithm is used which takes the following general form:

$$\sum_{c} \pi_c N(x\,;\, \mu_c, \sigma_c)$$

Here we have several mixture components that are indexed by c each of which is modelled by a Gaussian distribution. Each distribution has a mean (μ) and a covariance (σ) that represents the shape of the distribution while π represents the overall size of the distribution. Since we are working with multivariate data this formula changes slightly to the formula on the right. The only real difference here is that now we are using a covariance matrix to describe the distribution's variance as well as a vector mean to describe the mean. The EM algorithm has two parts, first the 'expectation part' which compares each data point under each distribution assigning it a responsibility factor that corresponds to the likelihood that the data point belongs in that cluster or distribution. The maximization part uses the responsibility factors from each point to better estimate each distribution's mean and corresponding variance. When an out of sample observation is to be assigned to a cluster, only the expectation part of the algorithm will be used.[iv] In my implementation I have used the "Mclust" R library in version 3.5.1.[v]

$$\mathcal{N}(\underline{x}\,;\, \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right\}$$

Results:



	diagnosis	
hclusters	2	4
1	448	49
2	8	192
3	2	0

Hierarchical clustering: Above we can see that the data is clustered into three distinct subgroups where the leftmost subgroup only represents a small fraction of the overall data. The most accurate result is obtained by cutting the tree at 3 clusters (horizontal dashed line) which gives us the result table above. The table shows that data is divided, 71%, 29% and .02% into the 3 clusters. These results are promising as they compare closely to the actual proportions of benign/malignant tumors mentioned above.

K-Means: The K-Means algorithm was run three separate times specifying 2, 3 and 4 clusters from which the three tables to the right were produced. Here we can see that K-Means produced the most accurate results when only 2 clusters were specified. The split of data was 65.3% and 34.6% between the two clusters. This matches closely with the actual diagnosis or results vector (65.5% and 34.5%).

							diagnosis	
	diagnosis						2	4
	diagnosis							
	2	4	1	10	199	2	447	1
1	9	233	2	0	34	3	0	33
2	449	8	3	448	8	4	9	101

Gaussian mixture model: The Gaussian mixture model chose a model with 4 clusters each with 23%, 43%, 19% and 15% of the data represented in each cluster. If the goal of this analysis was classification this would be a very poor result, however, because we are clustering, this distribution of data should lead the researchers to dive deeper into their study to figure out if there are further relationships between tumors that need to be studied.

	diagnosis	
	2	4
1	164	0
2	60	240
3	132	0
4	102	1

Conclusion: In the results section we found that the most accurate result was produced using the K-Means algorithm while specifying two clusters. However, this would only prove to be the *best* result had our goal of this analysis been classification. Since our analysis involved clustering, it would not be smart to rule out the results produced by the Hierarchical and Gaussian mixture models. Instead the results from these models could further indicate to the researchers that there are more than simply two subgroups within the data and that perhaps benign and malignant tumors can be broken down further to gain a better understanding of their health and diagnosis.

---

[i] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[ii] Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc..

[iii] Ibid.

[iv] Ibid.

[v] Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2017) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8/1, pp. 205-233