AMSI VACATION RESEARCH SCHOLARSHIPS 2019–20

EXPLORE THE MATHEMATICAL SCIENCES THIS SUMMER

Pitfalls of Machine Learning: Choosing Parameters and the Right Algorithm for the Data

Alexander Oakley
Supervised by Anya Reading
The University of Tasmania

Vacation Research Scholarships are funded jointly by the Department of Education and Training and the Australian Mathematical Sciences Institute.

Abstract

The amount of data in the world is growing exponentially. All this data may be useful human society, but one of the challenges of making use of this data is in dealing with the sheer volume of it. Clearly exponentially growing data cannot be comprehended by the human mind alone. The question is, can we use machine learning to find new information in these huge data sets. It seems that we can, but must be aware of the short comings of machine learning as we approach this problem. Here we look at several algorithms applied to the Iris Flower data set and find the Gaussian Mixture modelling works best.

1 Introduction

The amount of data in the world is growing at an astounding pace. According to Data [2013] 90% of the data that we now have was created in the last two years. What do we do with all this data? Is there a way for us to glean knowledge from it? Before we do that, can we turn this data into information? It seems like there should be something we can learn from it.

One of the challenges of making use of this data is in dealing with the sheer volume of it. Clearly exponentially growing data cannot be comprehended by a human mind alone. To make use of all the data available to us, we need some extra help. This where machine learning comes in. For most people, machine learning is the study of giving computer systems the ability to perform tasks without being explicitly instructed to do so. If we look 'underneath the hood' of this ability, we find a prerequisite skill is an ability to find patterns in data. Once patterns have been found, the machine can then turn data into information and then make informed decisions.

Humans do this constantly. Every time that we want to recognize an object that we are looking at, we need to first recognize its properties. To recognize its properties, we need to first recognize patterns. We do this mostly without being aware that we are doing so. For example, consider your ability to distinguish between a photo of a dog versus a photo of a cat. For most people, this task seems trivial. It is only when you are asked to describe what the difference is that you might realize that it is not so easy to pin down. Thankfully the pattern recognition abilities of you visual system to do not require you to be consciously aware of them.

If we can teach machines to find patterns in data, then can we use machine learning to automate knowledge discovery? [Wired magazine editor] believes that we can. He believes that we will simply be able to make inferences without making hypotheses, thus automating science. This sounds great, but is Chris Anderson being too optimistic? Can a data set be fed to the right set of machine learning

algorithms with the expectation that new knowledge will be produced, or is there more nuance to knowledge discovery than there appears to be? Silver [2012] says that "the numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning. Data-driven predictions can succeed - and they can fail".

As mentioned before, human beings are constantly finding patterns and new information in their surroundings. We transmute variations of light into images of faces, and variations in air pressure into sounds and then into speech. Sometimes when we do this it is not entirely clear if out perceptions match reality. There are many famous examples of people seeing faces where one should not be. The same is true of speech perception. A famous example is found in a song by the band, Led Zeppelin. There is a verse of their song, Stairway to Heaven that, when played backwards will sound like complete gibberish to most people. However, when it is played backwards alongside a visual aide that tells the listener what the lyrics *should* be, then gibberish becomes intelligible language. See the a presentation of this at `https://www.youtube.com/watch?v=7v57P1sfnHY`.

This example shows how data can be interpreted differently depending on the expectations of the interpreter. In an analogous way, the same is true for machine learning algorithms. Each algorithm only knows how to find the kinds of patterns that are recognisable to it. There are a plethora of different machine learning algorithms, each one with an ability to recognize a particular type of pattern. For a data analyst, the question is which algorithm suits my data.

Given this limitation, the question is; how do we choose the correct algorithm for data? Furthermore, how do we use the algorithm correctly? This report will discuss some of the ways of making that decision, and some the pitfalls of that one can run in to if applying machine learning without care.

Machine learning can be categorized into three broad types; reinforcement learning, supervised learning, and unsupervised learning. This report will focus on unsupervised machine learning.

2 Statement of Authorship

This work summarises the knowledge gained by Alexander Oakley during the summer of 2019/2020 while he worked on his Vacation Research Project that was funded by AMSI. Ross Turner and Anya Reading provided guidance. Most of the information in these pages came from SciKit-Learn and various blog posts and forums across the internet.

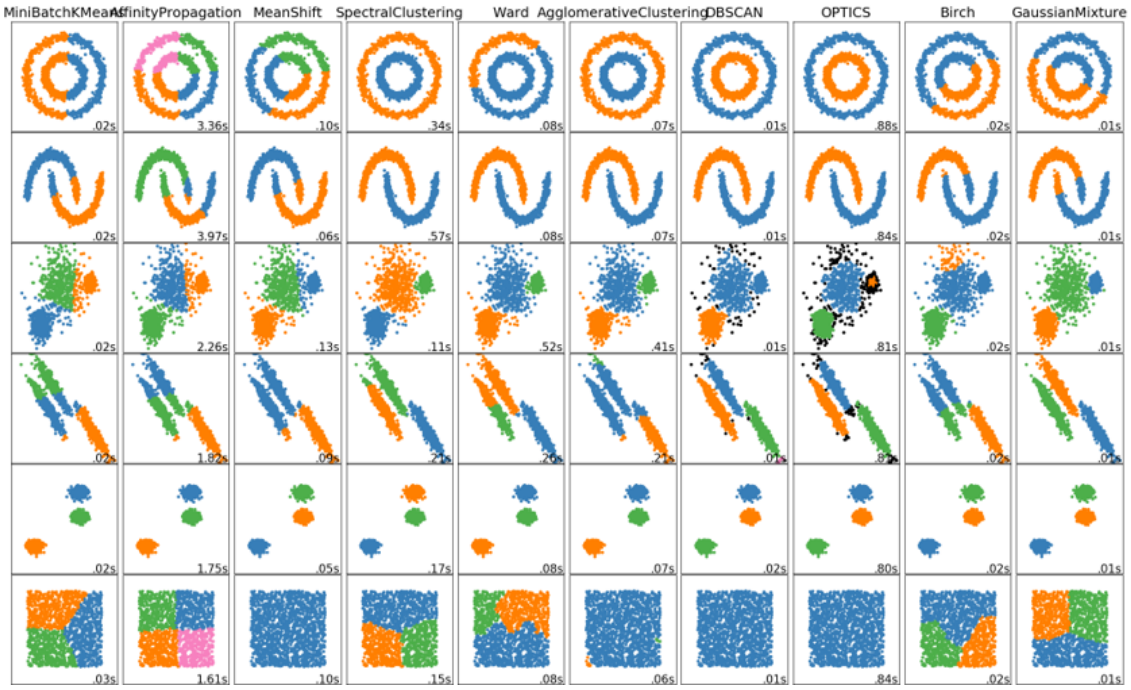

Figure 1: A collection of different two-dimensional data sets, and different ways to cluster the data points in them according to different unsupervised machine learning algorithms.

3 What is unsupervised machine learning?

Simply put, unsupervised machine learning is a process of putting "things" in to groups. These "things" are usually data points, and each data point has several attributes or features that are used to compare it to the other data points that are input to the algorithm. Each feature is usually a continuous variable, and together all the features constitute a feature space. Each unsupervised machine learning algorithm has its own method for defining groups. We call this process 'clustering', and its outputs 'clusters'.

As an example, consider a data set with a two dimensional feature space; height and weight (figure 2). This data set can partitioned in many different ways, clearly some ways are better than others, but different algorithms have different definitions of what a good output looks like (figure 3. This can be further illustrated by figure 1. Here we see range of different data sets (each row) and a variety of different clustering of those data sets from various unsupervised machine learning algorithms (each column).

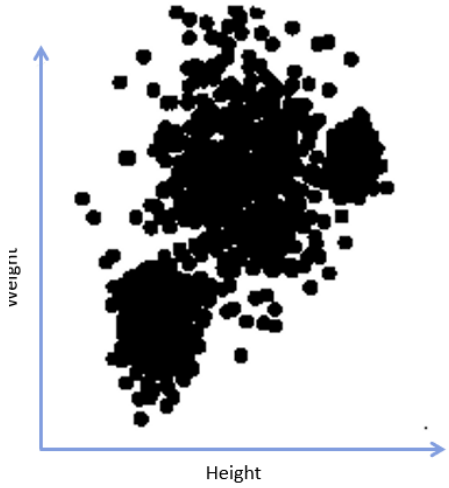

Figure 2: An imaginary 2 dimensional data set

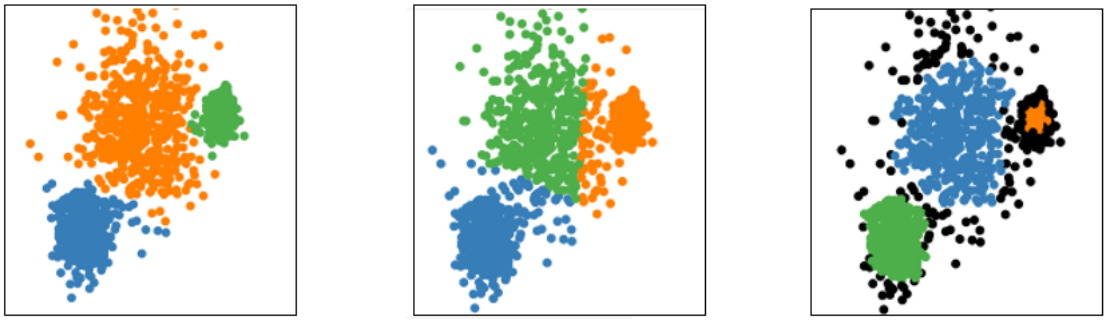

Figure 3: Three different ways to cluster the data shown in figure 2. Each color represents a different cluster. Each clustering is an output of a different unsupervised machine learning algorithm.

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Figure 4: Pseudo code for the K-Means algorithm.

4 K-Means

One of the most popular unsupervised machine learning algorithms is K-Means. K-Means works by grouping data points according to their proximity to 'centroids'. The first step in the algorithm is to place to centroids in the feature space. After that, in the second step is to assign each data point to its nearest centroid. The next step is to move each centroid to be at the centre of the data points that have been assigned to it. The process repeats from step two until either the change in centroid location is sufficiently small, or a fixed number of iterations have passed. The pseudo-code for the algorithm can be seen in figure 4 [scikitlearn, b]

The most important choice that needs to be made when using K-Means is, how many clusters do we want to try to find. This value is known as K and is the number of centroids that we initially place in to feature space at the beginning of the algorithm. To help choose K, we can use the silhouette score. Silhouette score gives us a measure of how 'good' the clustering is. This measure is a mixture of the cohesion within clusters (see SSE in equation 1 below) and the separation between clusters. To choose a good value for K, one might create several models, one for each value of K in a range, and then find the silhouette score for each of those models. The model with the highest silhouette score might be the best. Beware that the highest silhouette score is will not always find the 'correct' value for K.

An example of where the above method of choosing K fails to find the correct answer is seen when we apply K-Means to the famous Iris Flower data set. The un-clustered data set is visualized in figure 5. The silhouette score for K-Means models that use various values of K can be seen in figure 6. From this plot it seems that $K = 2$ is the appropriate choice. Using this value gives the plot seen in figure 7. This look pretty reasonable except that we know that the data comes from three different species of flower. In most cases, one might want to find three separate clusters, one for each species. A value of $K = 3$ gives the plot seen in figure **??**. This might seems valid until you see what the plot looks

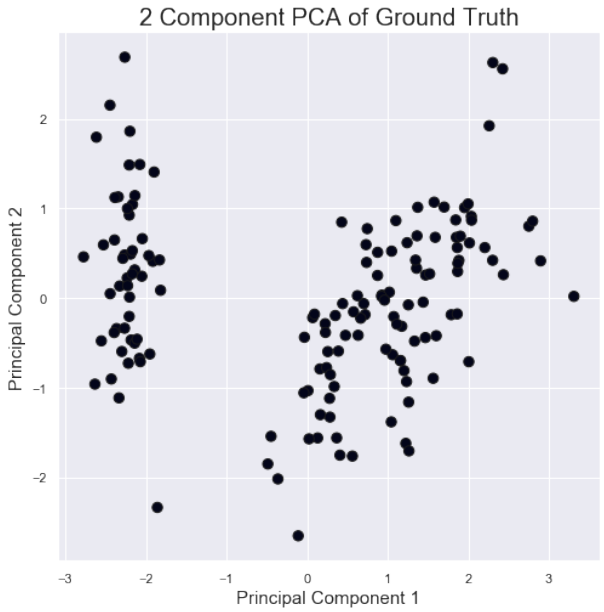

Figure 5: Here we see the Iris Flower data set visualized in two dimensional space using PCA.

like if we partition according to species (see figure 9).[scikitlearn, a]

K-Means is good at finding a 'centre-based' and compact clusters. Another way to describe these kinds of cluster is 'globular'. In mathematical terms, K-Means does clustering in a way that minimizes the squared sum of all distances of data points from their cluster centre (SSE) as per equation 1. In equation 1, K is the number of centroids, C_i is the i^{th} cluster, m_i is location of the i^{th} cluster in feature space, and x is a data point in feature space. Clearly, not all clusters are globular. Sometimes clusters can take on non-circular shapes, in which case SSE is not a relevant measure to minimize.

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist^2(m_i, x) \tag{1}$$

5 DBSCAN

Sometimes the clusters that we want to find are in shapes other than the globular ones the K-Means will find. Figure 10 shows data in that a human eye (and brain) would want to cluster in a way that is different to how K-Means would do it. Furthermore, the data seems to contain outliers that could be interpreted as noise. In most applications of clustering, we would want this noise to identified and not included in any of our clusters. DBSCAN does this.

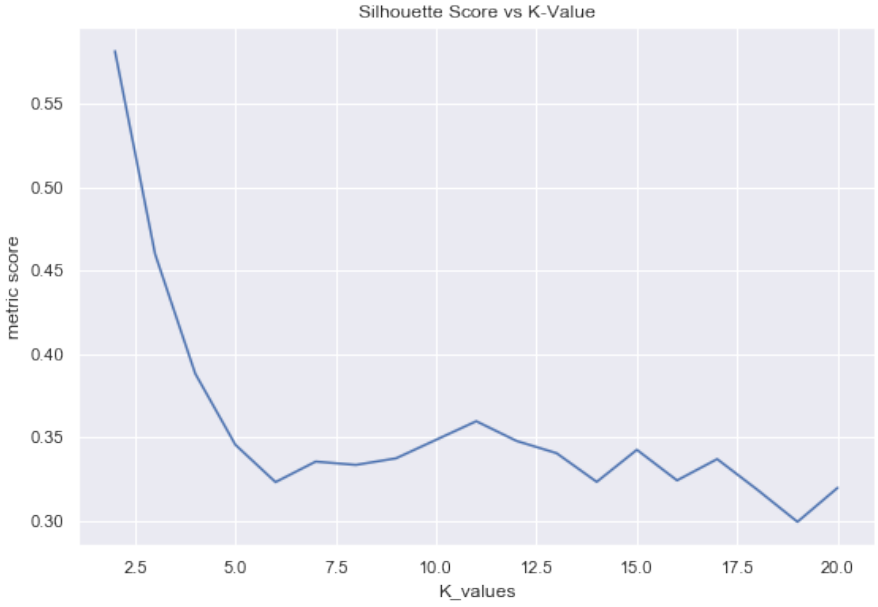

Figure 6: Here we see the silhouette score of the models output by K-Means for various values of K

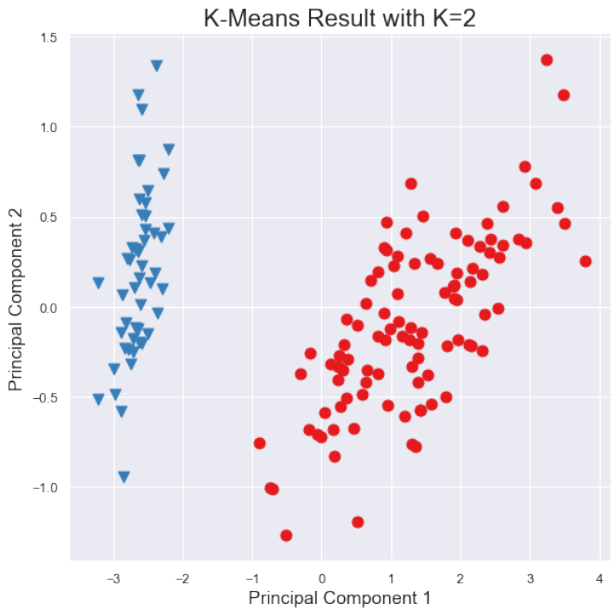

Figure 7: Here we see the Iris Flower clustered by K-Means with $K = 2$.

Figure 8: Here we see the Iris Flower clustered by K-Means with $K = 3$. 3 is the true number of classes, but K-Means has not found the true partition

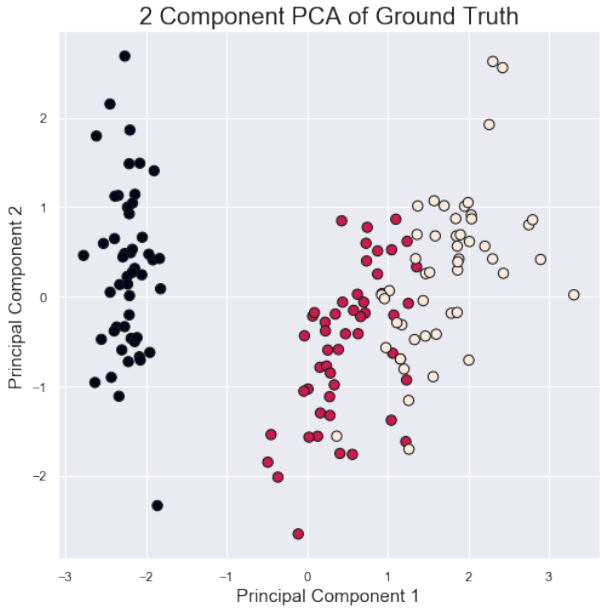

Figure 9: The true partition looks like this. Here we see three clusters, two of which are slightly overlapping and not globular. An appropriate clustering would have been Gaussian Mixture Models with Expectation Maximization which is good at finding elliptic clusters.

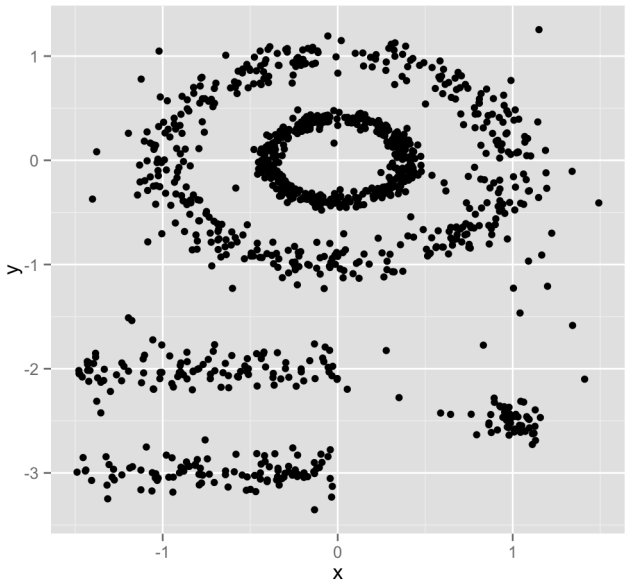

Figure 10: This is an example of data in which we might want to discover non-globular clusters.

6 Gaussian Mixture Modelling with Expectation Maximisation

Gaussian Mixture Modelling (GMM) with Expectation Maximisation (EM) makes the assumption that each variable comes from a Gaussian distribution. It takes as a parameter the number of separate distributions ($n-dist$) that each variable might have come from. The number of assumed distributions equates to the number clusters that GMM with EM will find.

To help choose a value for $n-dist$ we can use either the Akaike information criteria (AIC) or the Bayesian information criteria (BIC). If we use a consensus method, AIC and BIC together suggest a value of 3 for $n-dist$. Give this value, GMM with EM does a much better job of partitioning the Iris data set than K-Means does.

7 Conclusion

It is important to choose the right clustering algorithm. Here we have looked three different algorithms;; K-Means, DBSCAN, and GMM with EM. Although each algorithm has different abilities, GMM with EM has turned to be the most appropriate for our experimental data set.

The take home message here is that scientists and other investigators need to be aware of what sort of data they investigating and what the properties of it are.

Figure 11: Here we see the the AIC and BIC scores for various models that use different values of $n - dist$. Both indexes, the model with lowest score is preferred.

References

B. Data. for better or worse: 90% of world's data generated over last two years. *SCIENCE DAILY, May*, 22(3), 2013.

scikitlearn. Selecting the number of clusters with silhouette analysis on kmeans clustering¶, a. URL `https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html`.

scikitlearn. sklearn.cluster.kmeans, b. URL `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`.

N. Silver. *The signal and the noise: why so many predictions fail–but some don't.* Penguin, 2012.