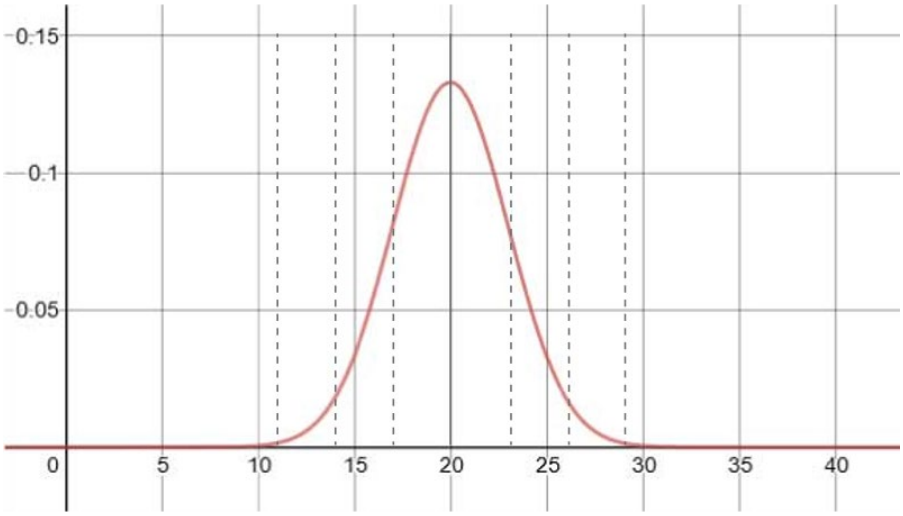
Tutorial Problems Week 4 – Solutions

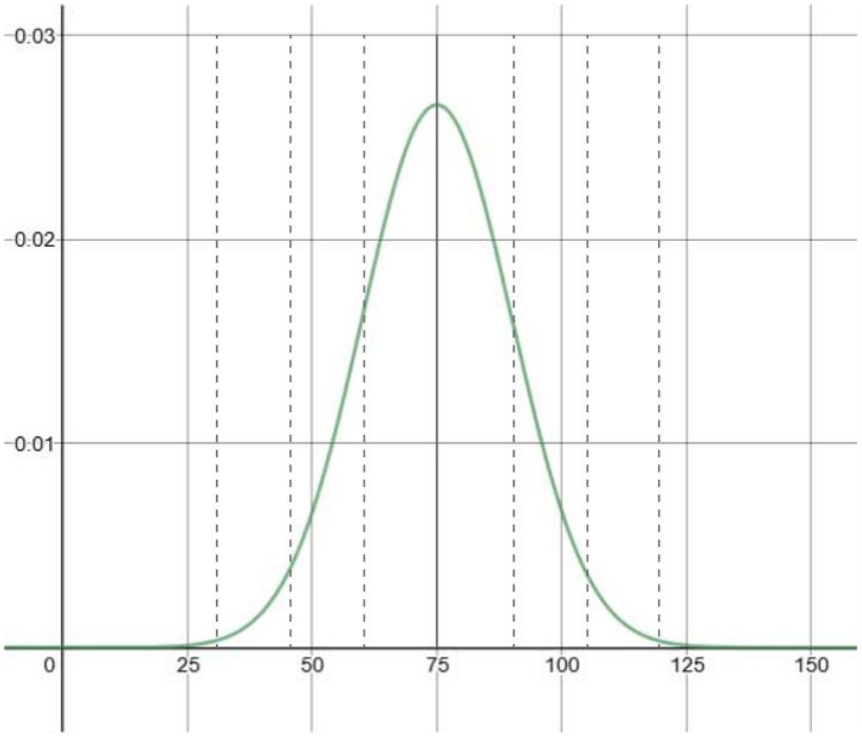
1 a) *Sketch the graph of a normal distribution with mean 20 and standard deviation 3. Remember that about 66% of the observations are within 1 standard deviation of the mean, about 95% are within 2 standard deviations and about 99% are within 3 standard deviations.*



When sketching the graph on a piece of paper there are three things to bear in mind,

1) The curve is the probability mass function of the distribution – The total area under the curve is 1.
2) The probability tops when x is equal to the mean of the distribution – in this case 20.
3) The probability would be very close to 0.0 beyond three standard deviations, as is also shown in the curve below.

b) *Sketch the graph of a normal distribution with mean 75 and standard deviation 15.*



Note the different scales between the two distributions.

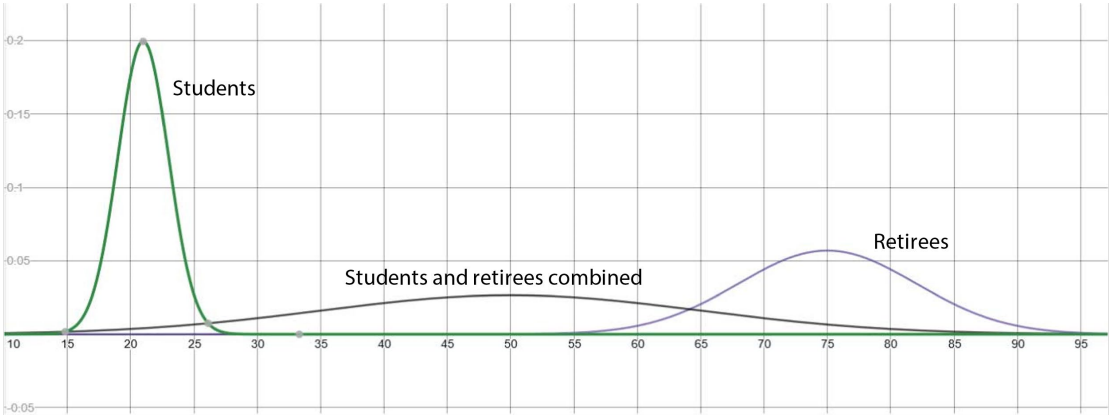
c) *Suppose you selected 100 RMIT students at random. What do you think the age distribution would be?*

The potential age range of RMIT students is quite large, the youngest student maybe as young as 14 years while the oldest may be in their 90s. If we limit our sample to undergraduate students, then the potential range of ages is going to be more limited. Most, (but not all) undergraduates would start their tertiary studies around the age of 18 or 19 years and complete their degree in three to four years (depending on the length of the degree). With this intuition, and the assumption that student ages are normally distributed (ie. they conform to a normal distribution), we could then assume an average (mean) age of about 21 years. A standard deviation of two years (for example), would mean that 66% of students would be aged between 19 and 23 years, 95% were aged 17 to 25 years (within two standard deviations), and 99% aged between 15 and 27 years (within three standard deviations). These estimations may or may not be correct, in order to check we would have to do the actual experiment ourselves!

d) *Suppose you selected 100 old age pensioners at random. What do you think that the age distribution would be?*

Again, we need to make some assumptions here. Lets replace the term "old age pensioners" with "retirees" and assume that the majority of people retire from the age of 65 (obviously some retire earlier and some later). We will also assume that few people live beyond the age of 100 years. As the probability of dying increases the older we get (excluding accidents) and not everyone retires at the same age we can also assume that the age distribution of retirees follows a normal distribution. Using these assumptions we can make an educated guess that the mean age of retirees is around 75 years with a standard deviation of maybe seven years.

e) *If you took the RMIT students and old age pensioners together, computed the mean and standard deviation and plotted the corresponding distribution, what would it look like? Sketch the distribution.*



The group of students and the group of pensions make two very different sets of data. Combining these two groups will result in a flat and wide bell curve.

f) *Explain how the above relates to clustering.*

The task of clustering involves grouping data instances so that those which are similar in some way are grouped together and there is clear separation between groups. We want to minimise the variance between items within each group and maximise the average distance (separation) between groups. A normal distribution with a small standard deviation indicates greater density of items and better similarity compared with a distribution with a larger standard deviation. Also, two normal distributions with very different means will have greater separation than distributions with similar means. Consequently, the task of clustering is to determine how many high thin distributions can be identified in the data and to ensure that they do not overlap too much.

2. a) *Give an interpretation of this output.*
The clustering is done on numeric data. The Instances fall into five roughly equal-sized groups. Cluster 0 consists of 28 Instances, the sepal length is around 4.7, the sepal width is around 3.1, the petal length is around 1.4 and petal width is around 0.19. Alternatively, one could say the sepal length ranges from about 4.29 to about 5.27 ……. (2 standard deviations).

Consider the sepallength. Cluster 0 is very well separated from other clusters. The closest one to it is 3 but the distance between their mean values is around 0.6, whereas the variance tells us at least 68% of the data within each cluster distribution (not real data points but an estimation) are very different (only around 16% overlapping). On the other hand, cluster 3 and 4 are very close though, their centroids are within a distance of around 0.26, while both their variances are comparable to this distance, which suggests that around half of their data points are within an overlapping range. Sepallength seems to be a good identifier for cluster 0. Similar analysis can be done on other attributes. When two clusters are similar in every row, they might be too close. For example, 1 and 2 are fairly similar and could possibly be merged into one cluster.

b) *There are three classes in the iris data. Why has the algorithm generated 5 clusters?*

N = -1. This means EM will find the number of clusters automatically based on cross-validation. For the given parameters and seed this turned out to be 5. For different runs a different number of clusters could be found. Also 3 classes in the data doesn't necessarily mean 3 clusters.

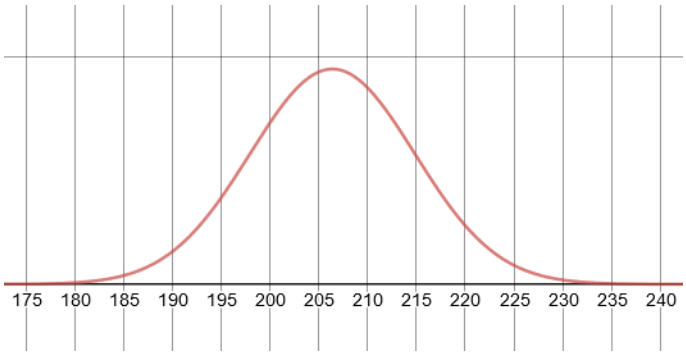
It is important to understand the difference between classes and clusters. Classes are based on labelled instances as determined by a domain expert. Clusters assume no labelling of instances and instances that are similar to each other are placed in the same cluster.

3. a) *What is the average and standard deviation for the basketball players?*

$$\text{Average (mean)} = \frac{213+210+214+200+195}{5} = 206.4$$

$$\text{Standard deviation} = \sqrt{\frac{(213-206.4)^2+(210-206.4)^2+(214-206.4)^2+(200-206.4)^2+(195-206.4)^2}{5-1}}$$
$$= 8.44$$

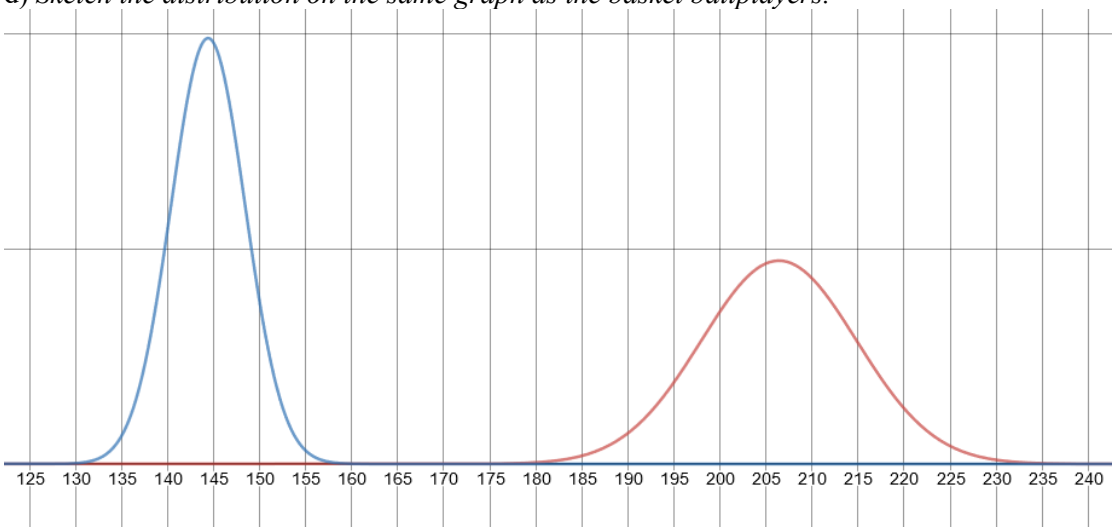
b) *Sketch the distribution.*



c) *What is the average and standard deviation for the jockeys?*

Average (mean) = 144.4
Standard deviation = 4.03

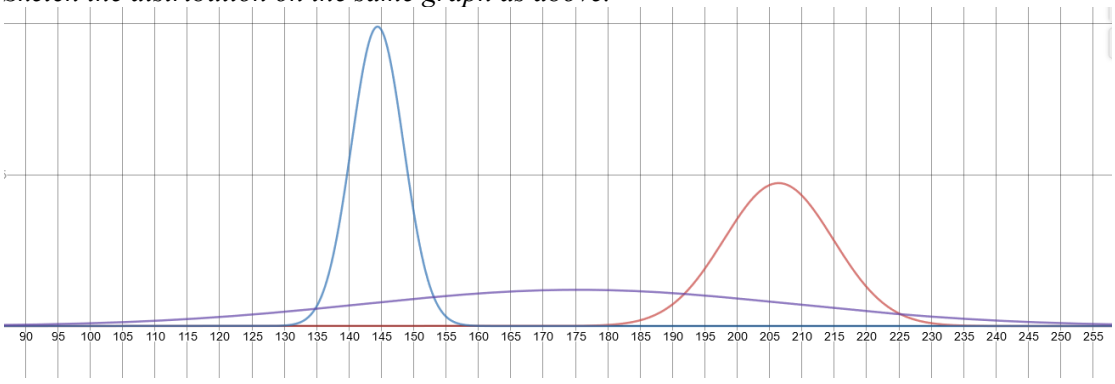
d) *Sketch the distribution on the same graph as the basket ballplayers.*



e) *What would be the average and standard deviation of all 10 individuals.*

Average (mean) = 175.4
Standard deviation = 33.26

Sketch the distribution on the same graph as above.



f) *Suppose that the EM algorithm was run on sports.arff. What would you expect the output to be.*

Two clusters reasonably well separated from each other.

4. a) *Starting with centroids (100 10) and (200, 20) work through three iterations of the K-means algorithm for K=2.*

Use Euclidean distance. (100,10) is initially in cluster 0, and (200,20) in cluster 1.
First iteration,

	DISTANCE TO…		
DATA	Centroid 0 (100, 10)	Centroid 1 (200, 20)	**Closest to centroid**
(213, 22)	114	13	**1**
(141, 35)	48	61	**0**
(210, 23)	111	10	**1**
(146, 31)	51	55	**0**
(214, 21)	115	14	**1**
(200, 24)	101	4	**1**
(195, 23)	96	6	**1**
(150, 30)	54	51	**1**
(145, 32)	50	56	**0**
(140, 31)	45	61	**0**

After the first iteration
Cluster 0: (141 35), (146 31), (145 32), (140 31)
Cluster 1: (213 22), (210 23), (214 21), (200 24), (195 23), (150 30)

New centroid locations:
Centroid of Cluster 0 = ((141+146+145+140)/4, (35+31+32+31)/4) = (143, 32.25)
Similarly, Centroid of Cluster 1 is (197, 23.83)

Now we repeat the process in the **second iteration**,
We shall notice that (150, 30) is closer to the new centroid of cluster 0 than cluster 1. The rest of the points remain unchanged.

After the second iteration
Cluster 0: (141 35), (146 31), (145 32), (140 31), (150, 30)
Cluster 1: (213 22), (210 23), (214 21), (200 24), (195 23)

New centroid locations:
Centroid of Cluster 0 = ((141+146+145+140+150)/5, (35+31+32+31+30)/5) = (144.4, 31.8)
Similarly, Centroid of Cluster 1 is (22.6, 206.4)

Third iteration,
Now it's fairly obvious that these data points will stay in the original clusters.

b) *What happens if the starting centroids are (177,28) and (178,29).*

In the first iteration the two centroids will move apart to the locations (144.4, 31.8) and (22.6, 206.4) respectively. This then forms the same clusters that were found in the second iteration of the question above (4a.), ie. Cluster 0: (141 35), (146 31), (145 32), (140 31), (150, 30) and

Cluster 1: (213 22), (210 23), (214 21), (200 24), (195 23). Therefore, using the centroids (177,28) and (178,29) as the starting points the algorithm will converge in one iteration of the algorithm.

c) *Is there a pair of starting seeds for which the algorithm won't converge as expected?*

Yes. If we want two clusters and one of our starting seeds is a long way from the data points and the other starting seed is close to the data points then the algorithm will stop before two clusters are identified. All data points would be allocated to the closer starting centroid and the more distance centroid would never be allocated any data points and would therefore never form a cluster or move closer to the data point locations. An example of such starting seeds for the example in 4(a) above could be (177, 28) and (5, 5), where the centroid at (5,5) would never form a cluster.

5. a) *Give an interpretation of this data.*

The following is just an EXAMPLE
1) Cluster 0, predominantly male, evenly split between MBC and BAPISCI, all around 20 years of age with a distinction average.
Cluster 2, mostly male, mostly MBC around 35 years of age with exceptionally very high marks.
Cluster 1, mostly female, mostly BAPPSCI around 20 years old with good marks.

b) *Are there any golden nuggets in this data?*

From the data it seems that male students do worse than female students in the course BAPPSCI. And students doing BAPPSCI are typically much younger than those doing MBC (older makes study harder?). A university sending ads to potential students probably should take these into consideration. And the staff team should analyse why male students do not perform as well as female in the course BAPPSCI. (There could be other nuggets. These are just examples.)