School of Computing and Information Systems
The University of Melbourne
COMP30027 MACHINE LEARNING (Semester 1, 2019)

Practical exercises: Week 8

Let's re-visit the three-class (8 or fewer, 9 or 10, 11 or more rings) *Abalone* task, but this time in an **unsupervised** context.

1. Use the method of k**-means** to cluster the *Abalone-3* data. (It's available from `sklearn.cluster.KMeans`)
   (a) Confirm that you understand the difference between the `fit()` and `predict()` methods in this context.
   (b) Write a function that calculates the **entropy** and **purity** of a k-means cluster, based on the true labels of *Abalone-3*.
   (c) Calculate the unsupervised evaluation metric known as **Sum of Squared Errors** on the resulting clusters.
   (d) k-means is typically re-run multiple times. In this case, it is controlled by the `n_init` parameter (which defaults to 10). Re-evaluate the resulting clusters for different values of `n_init` (especially 1 or 2, but perhaps also some larger values).
2. The most logical Expectation–Maximisation (EM) implementation within `scikit-learn` that is suitable for this data is a Gaussian Mixture model (`sklearn.mixture.GaussianMixture`; you can specify the number of clusters through the parameter `n_components`).
   (a) Contemplate what is happening in the "Expectation" step and in the "Maximisation" step for a **Gaussian** in this context.
   (b) This version of EM uses the centroids of k-means as a seed; what effect do you think this will have on its behaviour?
   (c) Compare the cluster evaluations of the EM clusters with the k-means clusters you calculated in the previous question. What do you observe, and why?