Part 2 Coursework 2 (20 marks)

Make sure you justify your answers with technical evidence - when in doubt, give details! Remember, any external material used **must be cited -** mark penalties will be applied.

1. Clustering (12 marks)

This part looks at clustering, a (unsupervised) learning technique not covered in-depth in class. Your goal is to understand the basics of a clustering algorithm, apply it to a sample dataset and draw conclusions about your findings. **Begin** by reading and studying the material about clustering on Blackboard ("ClusteringSlides CW2").

1.1. The results obtained from K-Means can vary significantly between runs due to what two facts about the initial centroid(s)? (1 mark)

1.2. For each of these two factors, provide a visual and technical explanation for why these can cause K-Means to get trapped in local minima. Remember, all material used must be cited.(4 marks)

1.3. Cluster the *baskball.arff* dataset (on Blackboard) using the *SimpleKMeans* method in Weka. Using a multiple set of values across the range for the number of clusters (K) from between 2 and 50, plot the sum of squared errors metric. State the trend observed, and provide an explanation for why this trend occurs. State and explain any other observations made. (3 marks)

1.4. Again using the *baskball.arff* dataset, generate a cluster model using k=3 clusters. What can you observe regarding players that have played for up to 25 minutes? Provide visualisations as appropriate. (Hint: visualise easily by right-clicking on the model and using Weka's visualisation). As a coach, how would you use clustering to help pick a team? Again, provide an explanation and some visualisations to support your answers, with reference to your previous cluster model. (2 marks)

1.5. What are some ways/methods of choosing k? You need to state and describe some methods for this, though an in-depth technical explanation is not required. (2 marks)

2. Association Rules: Mining a real-world dataset (8 marks)

Consider a real-world dataset, *vote.arff*, which gives the votes of 435 U.S. congressmen on 16 key issues gathered in the mid-1980s, and also includes their party affiliation as a binary attribute. This is a purely nominal dataset with some missing values (corresponding to abstentions). It is normally treated as a classification problem, the task being to predict party affiliation based on voting patterns. However, association-rule mining can also be applied to this data to seek interesting associations.

2.1. In Weka, run *Apriori* on this dataset with default settings. Comment on the rules that are generated. Discuss also their support, confidence and lift, showing you know how these are calculated, their role, and how to interpret the values. (5 marks)

2.2. It is interesting to see that none of the rules in the default output involve *Class* = *republican*. Why do you think that is? (3 marks)