AI Academy: Introduction to Data Mining
Week 2 Assessment

Please read and follow the instructions.

- **DUE DATE: This coming Sunday 11:45pm EST**
- **One Submission Per Person**

---

1. (9 points) [**Discretization**] Consider an attribute A_1 with the values shown in the box below.

   1. (3 points) Discretize the attribute A_1 by binning it into 4 equal-interval (i.e. equal-width) groups. Note that your bins should span from the sample minimum to the sample maximum.

-14	-12	-6	-5	-5	-2	-2	1	4	5	6	8	10	10	13	14

   2. (3 points) Discretize the attribute A_1 by binning it into 4 equal-frequency (i.e. equal-depth) groups. Note that your bins should span from the sample minimum to the sample maximum.

-14	-12	-6	-5	-5	-2	-2	1	4	5	6	8	10	10	13	14

   3. (3 points) Give an example of when it would make sense to discretize an attribute. Why would you expect the discretized attribute to be more useful for prediction or pattern recognition than the continuous attribute, even though it loses some data fidelity?

2. (11 points) [**Sampling**] Answer the following questions:

   (a) (5 points) A chess dataset contains records for 10,000 unique games, where 5,000 games result in a win for player white and 5,000 games result in a win for player black. Among the 10,000 unique games, 3% witness an en passant move and 11% have a castling move occur at least once on either the King's side or Queen's side. For simplicity, assume that the two events - en passant or castling - are mutually exclusive. Suppose we are developing a classifier in hopes of predicting the outcome of a chess game as it is being played. However, we are unable to use the entire data set due to computational limitations, and thus can only use a sample of the entire data set. Which sampling method would be appropriate and why? If we are sampling 2,000 games from the provided dataset, how many games should be selected from each group using your choice of sampling methods. Briefly justify your answer.

   (b) (6 points) Consider the following scenario: The data set originally had 10,127 games, but the 127 games that were omitted from consideration in the previous section did not result in a win for either player, but rather, a draw instead. Are the games that resulted in a draw, noise or outliers? Briefly justify your answer and describe the differences between noise and outliers.