Data Mining Qualifying exam.

Closed book, closed notes. Each problem is worth 20 points.

1. Use the candidate elimination algorithm to learn the concept of EnjoySport. You are given the following training set of 4 examples.

Sky	Temp	Humidity	Wind	Forecast	EnjoySport
S	W	N	H	S	No
C	C	H	L	C	Yes
S	W	H	L	C	Yes
R	C	N	M	C	No

Attributes can have the following values:

- Sky: S (Sunny), C (Cloudy), and R (Rainy)
- Temp: W (Warm) and C (Cold)
- Humidity: N (Normal) and H (High)
- Wind: L (Low), M (Moderate), and H (High)
- Forecast: S (Same) and C (Change)

A concept is represented as a 5-tuple of attrubute values ("?" stands for any value, "0" stands for no value). Show all your steps (Show sets S and G after the algorithm processes the first training example, the second training example, etc.).

2. Give two examples of distance measures (how to measure distance between the clusters) in hierarchical clustering and the implications of using them.

3. Use the Apriori algorithm to find all association rules with minimum coverage (support) 3 and minimum accuracy (confidence) 70% for the following data set. Show all your steps.

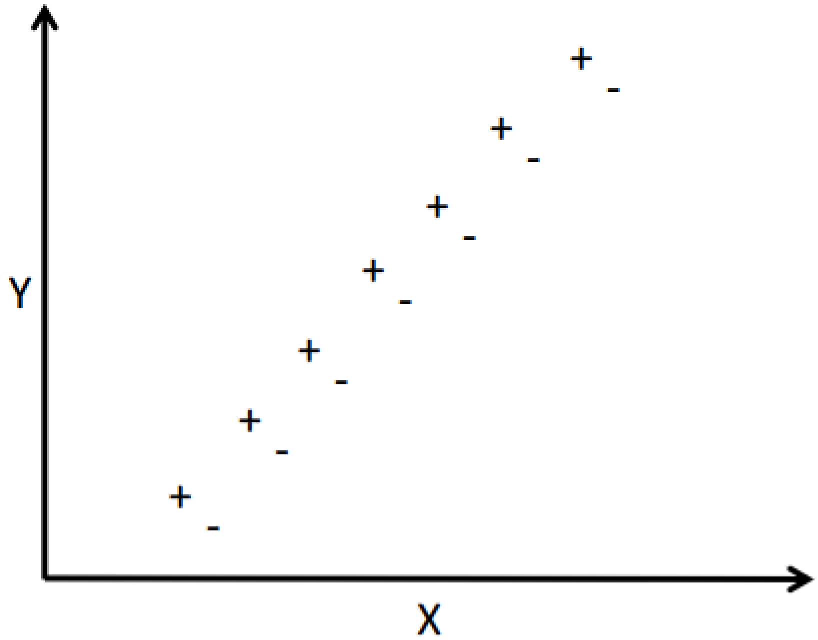
Transaction ID	Items
1	A, B, E
2	B, E, F
3	A, B, D, E
4	A, E, G
5	B, C, D, E
6	A, B, D, F, G
7	B, D, E

4. Discuss advantages and disadvantages of using prepruning and postpruning in decision trees.

5. Instance-Based Classification vs. Logistic Regression

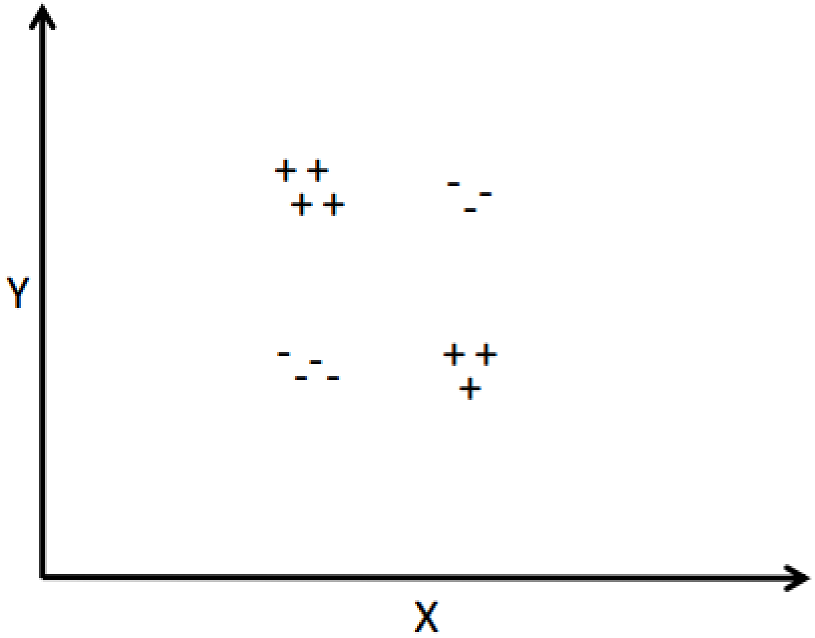
Suppose we have the data represented using two real-valued features (X and Y). Positive instances are marked with +, negative instances are marked with -. Distance between instances is measured using Euclidean distance. Suppose the data is randomly split into a training set (90%) and a test set (10%). Our goal is to train and evaluate a model.

(a) Suppose the data is the following:



Which classifier do you think would have a higher chance of doing well in terms of accuracy: k-Nearest-Neighbor (with k=1) or Logistic Regression? Why?

(b) Suppose the data is the following:



Which classifier do you think would have a higher chance of doing well in terms of accuracy: k-Nearest-Neighbor (with K=1) or Logistic Regression? Why?