

# Homework 11

CS 6243 – Spring 2005  
Tom Bylander, Instructor

assigned April 21, 2005  
due April 28, 2005

1. (64 pts.)  $k$ -means and EM clustering are available in Weka's explorer. Both of them produce measures of the goodness of the clustering. For the simplified Iris and Glass datasets, graph the values of these measures for  $k$ -means and EM clustering from  $k = 1$  to  $k = 5$  clusters for both datasets (4 graphs total). What value of  $k$  seems best in each case?
2. (36 pts.) For the weather dataset and a minimum support of 4, the following are the frequent 2-itemsets.

Item 1	Item 2	Support
outlook=overcast	class=pos	4
temperature=mild	humidity=high	4
temperature=mild	class=pos	4
temperature=cool	humidity=normal	4
humidity=high	windy=false	4
humidity=high	class=neg	4
humidity=normal	windy=false	4
humidity=normal	class=pos	6
windy=false	class=pos	6

In producing the frequent 3-itemsets:

- (a) What candidates are produced during the join step of the Apriori candidate generation algorithm?
- (b) What candidates are left after the prune step of the Apriori candidate generation algorithm?
- (c) What candidates have minimum support?