

## CMSC 726 Homework 5

Due Date: Thursday, April 27, at the start of class

1. For each of the following, find simple examples that illustrate the point. Specify the value of  $k$  that you are using.
  - (a) Construct a simple example (you can do this with 4 points) which shows that the cost of a clustering found by  $k$ -means by the cost of the optimal clustering (this ratio is called the approximation ratio for the approximation algorithm) can be arbitrarily high.
  - (b) Consider the  $k$ -medoids algorithm, where, instead of defining the cluster center as the centroid of the data points in the cluster, the cluster centers must be chosen from the data points. The goal of  $k$ -medoid algorithm is the same as  $k$ -means: minimize the total sum of distances from each data point to its cluster center. Construct a simple 1-d example where  $k$ -medians outputs a different clustering compared to  $k$ -means (starting with the same initial clustering).
2. Exercise 3.1 from ch. 3: The Bayesian Network Representation by Koller & Friedman (handed out in class).
3. Exercise 3.10.1 from ch. 3: The Bayesian Network Representation by Koller & Friedman. Extra Credit: 3.10.2 and 3.10.3.
4. In this question, you will explore the use of various clustering algorithms on your dataset from homework 1. If you prefer, you may use the dataset from your project, but please include a description of the dataset.
  - (a) We studied four general clustering algorithm families. Specific instances include:  $k$ -means, hierarchical agglomerative clustering, Gaussian mixture models and graph spectral clustering (for example using the top singular vectors or normalized cut). Choose TWO of these and implement them.
  - (b) Apply them to your dataset.
  - (c) 'Evaluate' your results. This is purposely vague, as the appropriate evaluation will depend on your dataset and choice of algorithms. However, please give sufficient thought here. It may be the case that you are limited to anecdotal analysis – this algorithm found this cluster of XXX, and that makes sense to me, while this other algorithm did not find this cluster. However the best answers here will include both a quantitative analysis and a qualitative analysis.
  - (d) Describe your results. Discuss any design choices you made. Include a printout of the main portions of your implementation.