Computational Linguistics Lecture Course, Summer 2016

Exercise 7: Clustering

You can earn up to 10 points on this exercise. 5 points is the lowest passing score.
You may submit individually or as a group of up to 3 people.
You may use any programming language you wish, but any submission that I cannot run on my computer without installing things must be presented to the class. (I like Python).

Please email your solution to `claytong@coli.uni-saarland.de` *by*
23:59 CEST on **June 30, 2016**. *Your name(s) should be present when I print the files you send!*

Download this two-dimensional dataset:
`http://www.coli.uni-saarland.de/courses/CL/2016/exercises/Exercise-8.dat`

Task 1

Implement the k-means algorithm for one dimensional data. Partition the dataset into 2 clusters using only the x values. Evaluate your partition visually. (2 points)

Partition the dataset into 2 clusters using only the y values. Is it better or worse? (1 point)

Task 2

Generalize your implementation to multi-dimensional data. Use squared Euclidian distance as your distance metric. Partition the dataset into 2 to 6 clusters using both dimensions. (2 points)

For best results, normalize the data in each dimension by *either*:

1. Dividing each dimension of squared distance by that dimension's variance (σ^2)
2. Converting each data point to a z-score, i.e. let x_i be the x-value of the ith data point and y_i be the y-value of the ith data point. Then,

$$z_i = (\frac{x_i - \mu_x}{\sigma_x}, \frac{y_i - \mu_y}{\sigma_y})$$

(1 point)

Task 3

Choose (a) or (b). Only one is required, but we will discuss both in class.

(a) Implement the Variance Ratio Criterion (VRC) as shown on slide 44.
How many clusters does VRC recommend for this dataset? (4 points)

(b) For general two-dimensional data, is there a $t_i = ax_i + by_i$ with suitably chosen a and b such that k-means clustering based on t_i is better than k-means clustering based on x_i or y_i alone? If so, prove it. If not, provide a counterexample. (4 points)