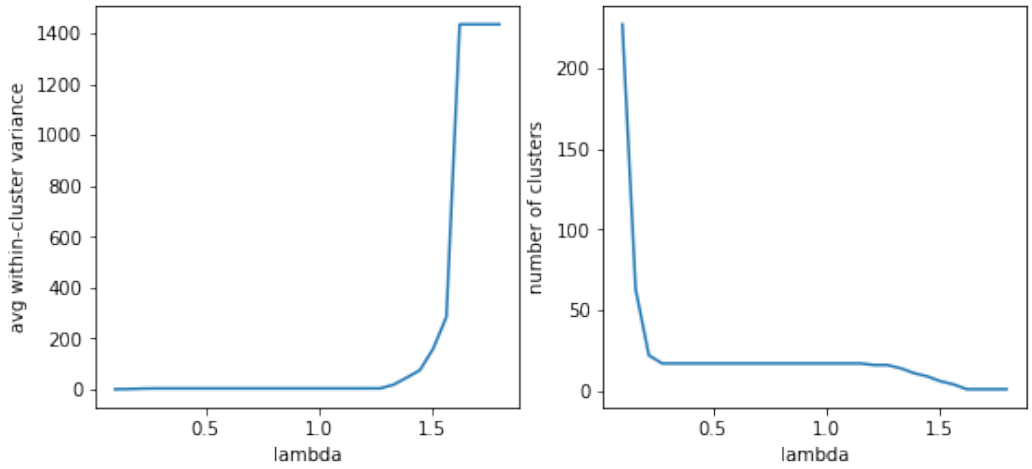
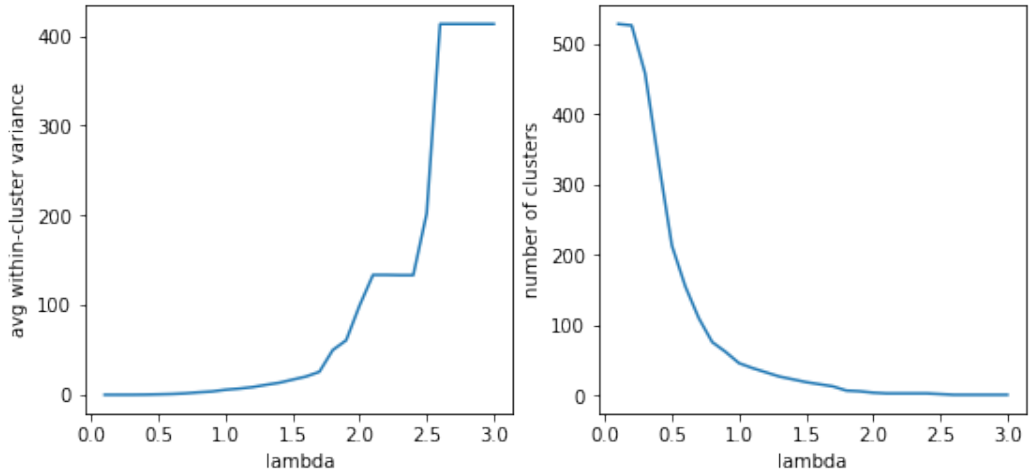
Yanbing Wang
11/18/19
K-means and clustering algorithm

Programming

Numerical data



Water treatment data



Discussion and results

The shape of the curves for both datasets look similar: there is a range for lambda to yield optimal results of the cluster. The lower bound is the bending point of the lambda vs. # clusters curve; the upper bound is the bending point of the lambda vs. variance curve. From this observation, we see that the optimal lambda range for the numerical dataset is 0.3-1.2, and for the water treatment data is 1.0-1.5.

Written

1) The total variance is a non-increasing function of k because:
   k-means converges because the within-cluster variance monotonically decreases or stays the same in each iteration. First, the variance decreases in the reassignment step since each vector is assigned to the closet centroid, so the distance it contributes to the variance decreases. Second, it decreases in the re-computation step because the new centroid is the vector for which the variance reaches its minimum.

   Analytically:

$$RSS_k(\mu_k) = \sum_{k=1}^{K} \sum_{x_i \in S_k} \|x_i - \mu_k\|_2^2$$

$$\frac{\partial RSS_k(\mu_k)}{\partial \mu_k} = \sum_{x_i \in S_k} 2(\mu_k - x_i)$$

   $\mu_k = \frac{1}{n_k} \sum_{x_i \in S_k} x_i$, where nk is the number of points in the cluster k.
   Thus, we minimize RSS_k when the old centroid is replaced with the new centroid. RSS, the sum of the RSS_k, must then also decrease during recomputation.

2) See attachment

b) $x \to \varphi(x)$

$$\sum_{k=1}^{K} \sum_{x_i \in S_k} \| \varphi(x_i) - a_k \|_2^2$$

$$= \sum \sum (\varphi(x_i) - a_k)^T (\varphi(x_i) - a_k)$$

$$= \sum \sum \varphi(x_i)^T \varphi(x_i) + a_k^T a_k - 2\varphi(x_i)^T a_k$$

$$a_k^T a_k = \frac{1}{|S_k|^2} (\sum \varphi(x_i))^T (\sum \varphi(x_i))$$

$$= \frac{1}{|S_k|^2} \sum_{i,j \in S_k} K(x_i, x_j)$$

$$\varphi(x_i)^T a_k = \varphi(x_i)^T \frac{1}{|S_k|} \sum_j \varphi(x_j) = \frac{1}{|S_k|} \sum_{j \in S_k} K(x_i, x_j)$$

∴ the original objective fcn. becomes:

$$\min_{S=\{S_1 \cdots S_K\}} \sum_{k=1}^{K} \sum_{x_i \in S_k} \left[K(x_i, x_i) + \frac{1}{|S_k|^2} \sum_{i,j \in S_k} K(x_i, x_j) + \frac{2}{|S_k|} \sum_{i,j \in S_k} K(x_i, x_j) \right]$$

$K(x_i, x_i)$: ↓ do not change

$\frac{1}{|S_k|^2} \sum_{i,j \in S_k} K(x_i, x_j) + \frac{2}{|S_k|} \sum_{i,j \in S_k} K(x_i, x_j)$: fcn. of $K(x_i, x_j)$, measure of within-cluster variance.