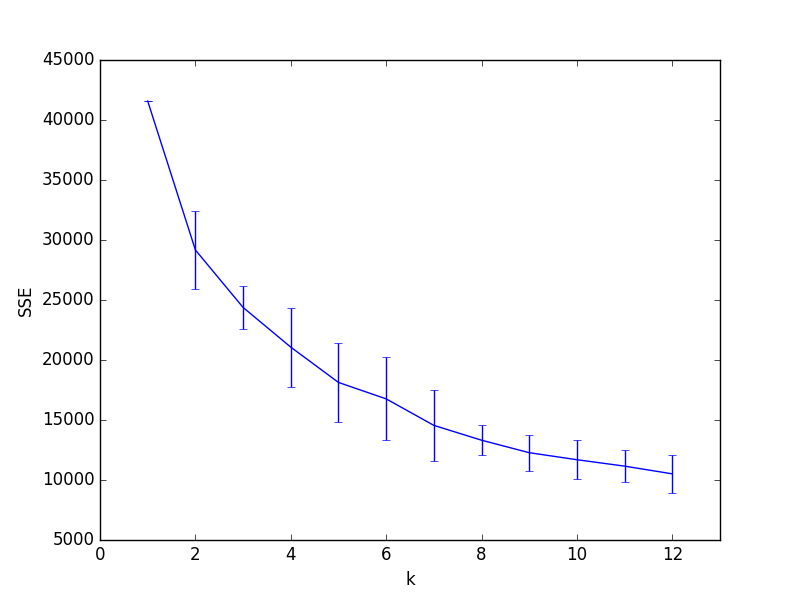
CS6220-01 Data Mining Techniques – Fall 2016

Assignment 3

Qian Cao

**Question 5-Results & Answers**

****

k= 1 uk= 41580.0 uk-2std= 41580.0 uk+2td= 41580.0

k= 2 uk= 29171.0466007 uk-2std= 25897.9022197 uk+2td= 32444.1909818

k= 3 uk= 24380.3052221 uk-2std= 22586.796765 uk+2td= 26173.8136793

k= 4 uk= 21072.7135239 uk-2std= 17778.891225 uk+2td= 24366.5358228

k= 5 uk= 18131.1066598 uk-2std= 14805.6062263 uk+2td= 21456.6070933

k= 6 uk= 16760.9544458 uk-2std= 13294.2670538 uk+2td= 20227.6418378

k= 7 uk= 14552.4712704 uk-2std= 11622.143152 uk+2td= 17482.7993889

k= 8 uk= 13316.8626499 uk-2std= 12047.6164637 uk+2td= 14586.1088362

k= 9 uk= 12275.703526 uk-2std= 10784.466063 uk+2td= 13766.940989

k= 10 uk= 11690.3339587 uk-2std= 10088.6919828 uk+2td= 13291.9759345

k= 11 uk= 11155.269765 uk-2std= 9840.14060201 uk+2td= 12470.3989279

k= 12 uk= 10513.030699 uk-2std= 8957.82152612 uk+2td= 12068.2398719

* **Question: As k increases and approaches the total number of examples N, what value does the SSE approach? What problems does this cause in terms of using SSE to choose an optimal k?**

Answer: The SSE approaches zero as k increases and approaches the total number of examples N, which means there is no squared errors. In terms of using SSE to choose k, since we want SSE to be the smallest, the best k will locate at k = N, where SSE = 0. Obviously, it not the best choice for k, which is too large for a huge dataset and does not make any sense for dividing N examples into N clusters.

* **Question: Can you suggest another measure of cluster compactness and separation that might be more useful than SSE?**

Answer: Absolute error criterion may be more useful than SSE.