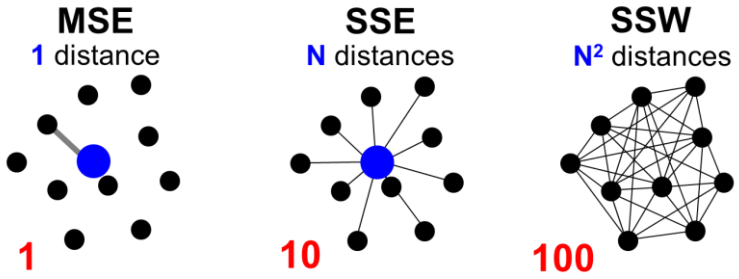
Clustering Methods

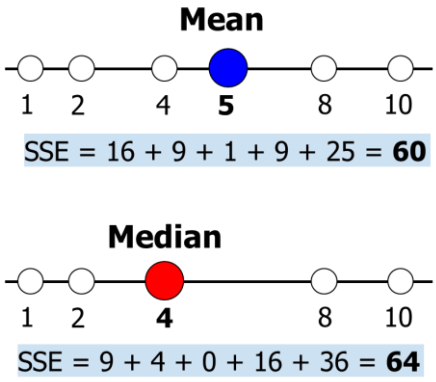
Exercises 1/7

1. There are three alternative ways to measure goodness of a cluster: MSE, SSE and SSW as shown below. Which one would you choose in the following situation and why?

   a. You want to avoid very small cluster sizes.
   b. You want clusters with as small variance as possible.
   c. You have text data and use *edit distance* to measure distance between data points.
   d. You want to separate outlier points from normal clusters.



2. K-means calculates centroid as the mean of each attribute value in the cluster whereas another algorithm, K-median, would use the median of the attribute values (see below). Prove that mean value is the optimal choice for minimizing SSE.



3. How would you evaluate the balance of cluster sizes? Define at least one measure and explain the reasoning. What value would it give if all cluster sizes were equal? What is the worst case for this measure.

4. Balanced k-means formulates the point-to-cluster assignment as graph problem by creating n/k slots in each cluster and then finding the optimal pairing of data points and the slots. How many edges are there? Can you think of another way to define the assignment as a graph problem without so many edges. What algorithmic problem is this?

Bonus task:

5. Find details of Hungarian algorithm anywhere on web. Study the details and write overall description of the main steps of the algorithm. Explain where the $O(n^3)$ time complexity comes from.