Clustering

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Outline

1. Clustering
2. Distance Measures for Clustering
3. K-Means
4. Bath Soap Example
5. Choosing the # Clusters
6. Fun Example

Public Utilities Data

Clustering; also called **data segmentation**:

- Goal: Partition the dataset into clusters (groups) such that
  - Data within each cluster are similar to one another
  - Data in two different clusters are dissimilar
- Applications:
  -
  -
  -
  -

Example: Public Utilities Data

Utilities example:

- Have data on public utilities (Shmueli, Patel, and Bruce, 2007)
- Wish to group based on financial factors
- Used for e.g. a study on the impact of deregulation
- Could do a detailed cost/impact analysis for a "typical" utility in each group
- Scale up to estimate impact for all utilities
  - Saves money relative to doing detailed cost/impact analysis for every utility

Utilities Data

Data on public utilities (Shmueli, Patel, and Bruce, 2007):

	Company	Fixed.charge	RoR	Cost	Load.factor	Demand.growth	Sales	Nuclear	Fuel.Cost
1	Arizona	1.06	9.2	151	54.4	1.6	9077	0.0	0.628
2	Boston	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
3	Central	1.43	15.4	113	53.0	3.4	9212	0.0	1.058
4	Commonwealth	1.02	11.2	168	56.0	0.3	6423	34.3	0.700
5	NY	1.49	8.8	192	51.2	1.0	3300	15.6	2.044
6	Florida	1.32	13.5	111	60.0	-2.2	11127	22.5	1.241
7	Hawaiian	1.22	12.2	175	67.6	2.2	7642	0.0	1.652
8	Idaho	1.10	9.2	245	57.0	3.3	13082	0.0	0.309
9	Kentucky	1.34	13.0	168	60.4	7.2	8406	0.0	0.862
10	Madison	1.12	12.4	197	53.0	2.7	6455	39.2	0.623
11	Nevada	0.75	7.5	173	51.5	6.5	17441	0.0	0.768
12	New England	1.13	10.9	178	62.0	3.7	6154	0.0	1.897
13	Northern	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
14	Oklahoma	1.09	12.0	96	49.8	1.4	9673	0.0	0.588
15	Pacific	0.96	7.6	164	62.2	-0.1	6468	0.9	1.400
16	Puget	1.16	9.9	252	56.0	9.2	15991	0.0	0.620
17	San Diego	0.76	6.4	136	61.9	9.0	5714	8.3	1.920
18	Southern	1.05	12.6	150	56.7	2.7	10140	0.0	1.108
19	Texas	1.16	11.7	104	54.0	-2.1	13507	0.0	0.636
20	Wisconsin	1.20	11.8	148	59.9	3.5	7287	41.1	0.702
21	United	1.04	8.6	204	61.0	3.5	6650	0.0	2.116
22	Virginia	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

Utilities Data

- 8 operational variables:

  Fixed.charge: Fixed-charge covering ratio (income/debt)
  RoR: rate of return on capital
  Cost: cost per kilowatt capacity
  Load.factor
  Demand.growth
  Sales: Kilowatthour use per year
  Nuclear: % nuclear
  Fuel.cost: Total fuel costs

Clustering

Want to partition dataset into clusters such that

- Observations within each cluster are similar to each other
- Observations in two different clusters are dissimilar

So the key to clustering will be defining an appropriate distance measure!

2 types of clustering methods:

- **Distance-based clustering:** uses an explicit distance measure to group observations
- **Model-based clustering:** fits a statistical model that captures groups in the data; no explicit distance measure

Distance Measures

Need a measure of distance $D(x, x')$ between any two observations x and x' (p-dimensional vectors).

- Example for continuous variables:
- Example for categorical variables:
- Ordinal variables are generally encoded as integers then treated as continuous.

Distance Measures

Options for continuous variables:

(a) $\sum_{j=1}^{p} |x_j - x'_j|$

(b) $\sum_{j=1}^{p} (x_j - x'_j)^2$

(c) $\sum_{j=1}^{p} |x_j - x'_j|^{1/2}$

Properties:

Standardization

Why it's a good idea to standardize continuous variables before distance-based clustering:

Because different variables have different units and variances, and without standardization the high-variance variables dominate the (e.g. Euclidean) distance.

Example:

Distance Measures

- Lesson: explore the data using EDA as much as possible, and be careful about the choice of distance measure

- Typically standardization is a good practice

Distance-Based Clustering Methods

Distance-Based Clustering Methods

- There is a large assortment of such methods
- Out of distance-based clustering methods, the choice of distance measure may be the most important choice, **not** the choice of method

K-Means Algorithm

NOTATION:

$i = 1, \ldots, n$ labels the observations (objects) to be clustered

$k = 1, \ldots, K < n$ labels the group (cluster)

- K must be chosen. We'll talk about that later.

$C(\cdot)$ is an "encoder", i.e. $C(i) = k$ if observation i is assigned to cluster k.

K-Means Algorithm

K-Means Algorithm:

- Attempts to choose an encoder C that minimizes the within-cluster variability:

$$W(C) = \sum_{k=1}^{K} \sum_{i:C(i)=k} \|x_i - \bar{x}_k\|^2$$

where $\bar{x}_k$ is the sample mean for the kth cluster, $\bar{x}_k = \frac{\sum_{i:C(i)=k} x_i}{\sum_{i:C(i)=k} 1}$

  - Idea: observations in the same cluster should be close to each other

Note:

K-Means Algorithm

K-Means:

- An iterative “greedy descent” algorithm
- Only appropriate if all variables are continuous
- Uses distance measure

$$D(x, x') =$$

- Can standardize the variables first

K-Means Algorithm

Training (learning C from training data)

Start with some assignment C of observations to clusters.

Iterate back and forth between

1. Define the mean of cluster k to be

2. Assign object i to the cluster it is closest to, that is, object i is assigned to cluster k if

Stop when C does not change.

K-Means Algorithm

One can also start with arbitrary cluster means $m_1, m_2, \ldots, m_K$ then perform (2), then (1), then (2), etc.

In either case, one should

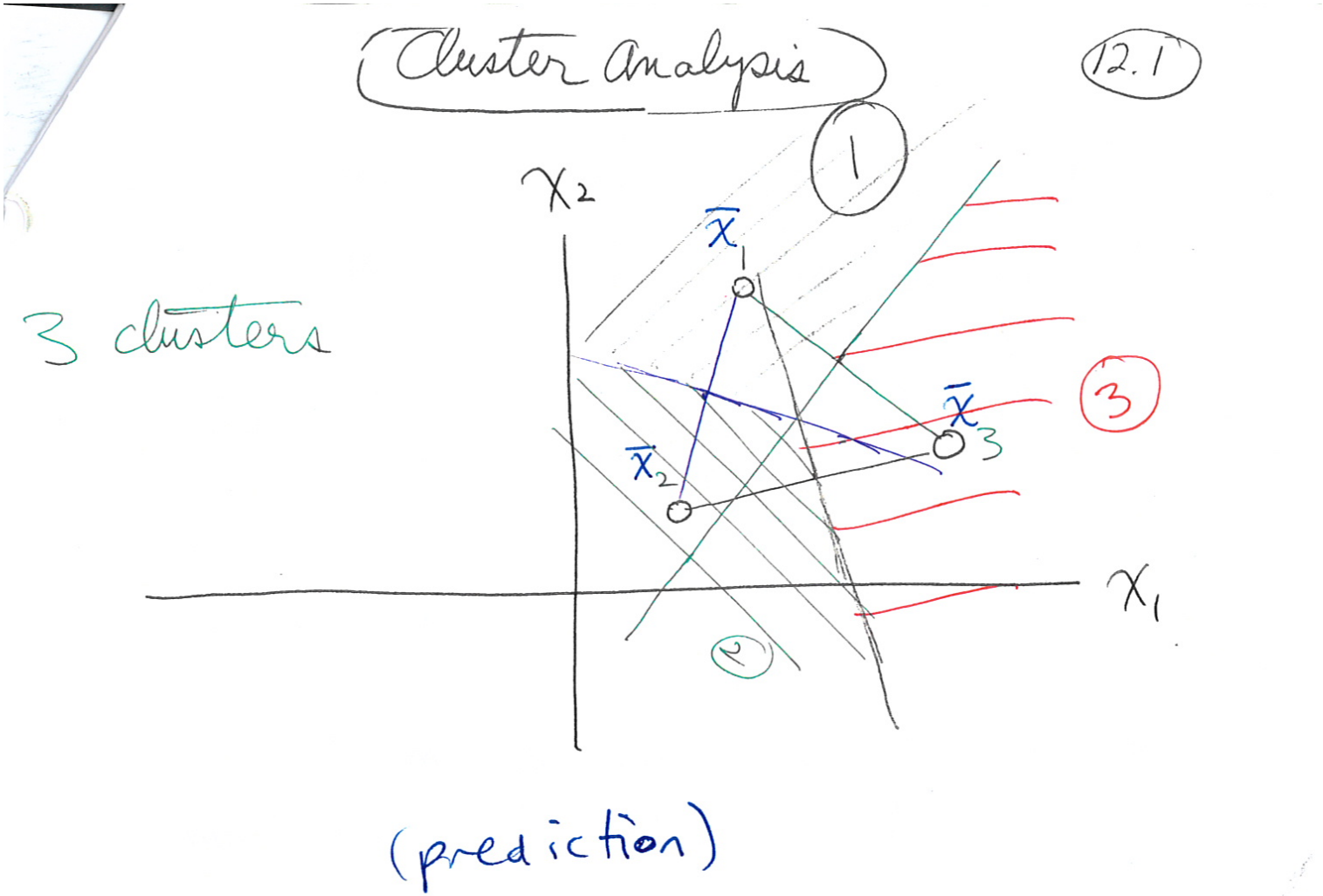
K-Means Algorithm

Prediction:

The encoder C that is ultimately selected can be used to assign new observations to clusters. Assign the new observation to the cluster with the closest mean m_k.

This can be viewed as partitioning the "X-space" into regions according to what the cluster assignment would be for a new observation.

K-Means Algorithm



K-Means Algorithm

- In each step of the k-means algorithm the objective function

$$W(C) = \sum_{k=1}^{K} \sum_{i:C(i)=k} \|x_i - m_k\|^2$$

  decreases.

- So k-means converges to a local minimum of $W(C)$. Not a global minimum!

K-Means Algorithm

Why does $W(C)$ always decrease?

Bath Soap

Bath Soap Example (Shmueli et al.):

- A maker of bath soaps in India is interested in segmenting its customer base in order to do more effective marketing
- It has data on customer demographics and purchase history
- The company can develop 2-5 different promotional approaches.

Bath Soap

- This is clearly a clustering problem.
- The # of clusters is primarily determined by the requirements of the company

Bath Soap

	A	J	K	L	M	N	O	P	Q	R
1	Household			Purchase Summary						
2										
3	**Member id**	**CS**	**Affluence Index**	**No. of Brands**	**Brand Runs**	**Total Volume**	**No. of Trans**	**Value**	**Trans / Brand Runs**	**Vol/Tran**
4	1010010	1	2	3	17	8025	24	818.00	1.41	334.38
5	1010020	1	19	5	25	13975	40	1681.50	1.60	349.38
6	1014020	1	23	5	37	23100	63	1950.00	1.70	366.67
7	1014030	0	0	2	4	1500	4	114.00	1.00	375.00
8	1014190	1	10	3	6	8300	13	591.00	2.17	638.46
9	1017020	1	13	3	26	18175	41	1705.50	1.58	443.29
10	1017110	1	11	4	17	9950	26	1007.50	1.53	382.69
11	1017160	0	0	3	8	9300	25	569.50	3.13	372.00
12	1017360	1	17	2	12	26490	27	3113.50	2.25	981.11
13	1017460	1	6	4	13	7455	18	990.50	1.38	414.17
14	1017490	1	30	4	24	16275	42	1555.50	1.75	387.50
15	1020070	1	18	4	21	13875	61	1547.75	2.90	227.46
16	1020210	2	8	3	9	20675	26	1538.00	2.89	795.19
17	1024050	1	10	4	10	15450	28	1275.50	2.80	551.79
18	1024100	1	15	2	8	9150	25	917.00	3.13	366.00
19	1024120	2	13	4	19	16050	50	1518.00	2.63	321.00
20	1024220	0	11	7	28	19150	38	1914.00	1.36	503.95
21	1024400	0	4	4	13	14625	31	1395.00	2.38	471.77
22	1024630	1	14	3	12	14400	21	1308.50	1.75	685.71

Bath Soap

Purchase History Variables:

- Number of brands purchased
- “Brand runs”
- Total volume
- Number of transactions
- Total value
- Average volume per transaction
- Average price of purchase
- Percent of volume purchased within promotion
- Percent of volume purchased in highest price category
- ...

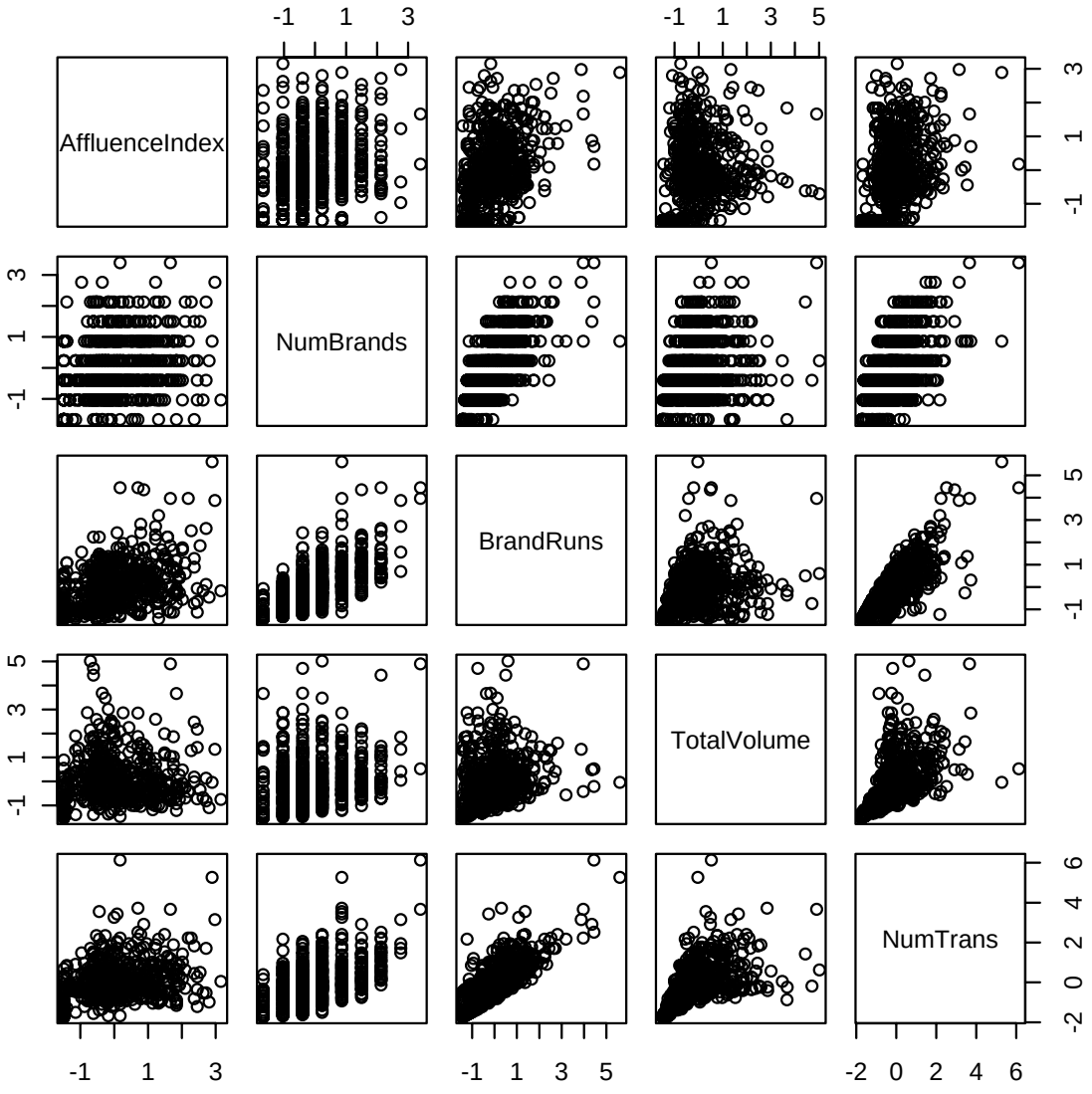
Bath Soap

Measuring brand loyalty

- Many of these variables are intended to measure brand loyalty, a major interest of the company
- # different brands: higher means poor brand loyalty
- Percent of volume purchased in a single brand: higher means good brand loyalty

Bath Soap

Pairwise scatterplots of several variables:



Bath Soap

- I standardized the continuous variables, and used them to do k-means clustering
- I did NOT use the categorical demographic variables (e.g. language spoken) since k-means can't handle these

Bath Soap

K-means results with 2 clusters:

- 600 customers total
- 72 put in one cluster, 528 put in other

K-means with 3 clusters also put >500 customers in a single cluster

Bath Soap

K-means results with 4 clusters:

- More balanced cluster sizes (biggest is 200)
- Can look at the cluster centroids to help interpret the clusters
- Remember that the variables have been standardized, so the values represend # SDs above / below the mean
- To get the following output in R I just typed “soapClust” at the command line, where “soapClust” is the object returned by the call to “kmeans” function

Bath Soap

```
Cluster means:
  AffluenceIndex  NumBrands  BrandRuns  TotalVolume   NumTrans
1      0.2103270 -0.2725564 -0.1464591 -0.635757690 -0.2983335 -0.
2     -0.7855210 -0.5658063 -0.7868276  0.130221890 -0.4153865 -0.
3     -0.2320878 -0.3871225 -0.3946937  0.007046329 -0.3885536 -0.
4      0.3680956  0.7617435  0.7640406  0.363832528  0.7274045  0.
  TransOverBrandRuns VolOverTran    AvgPrice PurVolNoPromoPerc Pur
1        -0.18970418  -0.5281348  1.25067498       -0.34596731
2         1.03963946   0.5458295 -1.32810105        0.18944829
3        -0.08336143   0.3567326 -0.32456753        0.17944255
4        -0.15667295  -0.2000610 -0.03141205       -0.01824366
  PurVolOtherPromoPerc       BrCd1     BrCd55     BrCd272     BrCd2
1           0.07636960 -0.3494306 -0.4634947  0.63418870 -0.206285
2           0.22650962 -0.5761328  2.4965133 -0.30358436 -0.247411
3          -0.03766242  0.6765137 -0.3268173 -0.26072548  0.277241
4          -0.09182635 -0.2426077 -0.2450074 -0.04971722 -0.054143
       BrCd24    BrCd481    BrCd352       BrCd5      Others       P
1  0.58964968 -0.1605765 -0.1276905  0.35233970  0.42924154  1.310
2 -0.21772937 -0.2488248 -0.2702091 -0.16434001 -1.25438976 -0.802
3 -0.21895589 -0.1799151 -0.2207571 -0.17127281 -0.03548219 -0.494
4 -0.09215415  0.3712818  0.3978638 -0.00284705  0.19318799 -0.085
      PrCat2     PrCat3     PrCat4   PropCat5   PropCat6   PropCat
1 -0.5705789 -0.4840490 -0.3155730 -0.4156053  0.1578613  0.154601
```

Bath Soap

What cluster is this:

- Affluent
- High volume of purchases
- Large # transactions
- High value of purchases
- High # of brands
- Tends to purchase items in brands 481 and 352

Pick Answer:

A. Cluster 1

B. Cluster 2

C. Cluster 3

D. Cluster 4

Bath Soap

What cluster is this?

- Low affluence
- Low average price
- Low total value
- Low # transactions
- Low # brands
- Moderate total volume

Pick Answer:

A. Cluster 1

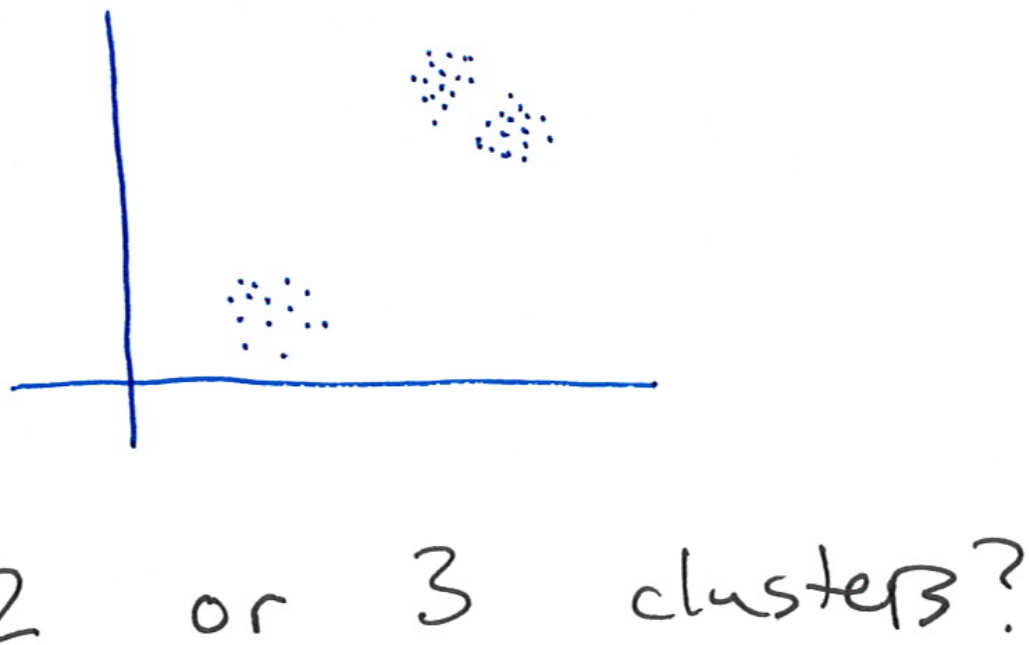
B. Cluster 2

C. Cluster 3

D. Cluster 4

Choosing the # Clusters

- How do we choose K, the number of clusters?
- The question of what is the "true" number of clusters is not well-defined:



Choosing the # Clusters

But how many clusters should we use? Several answers:

1.

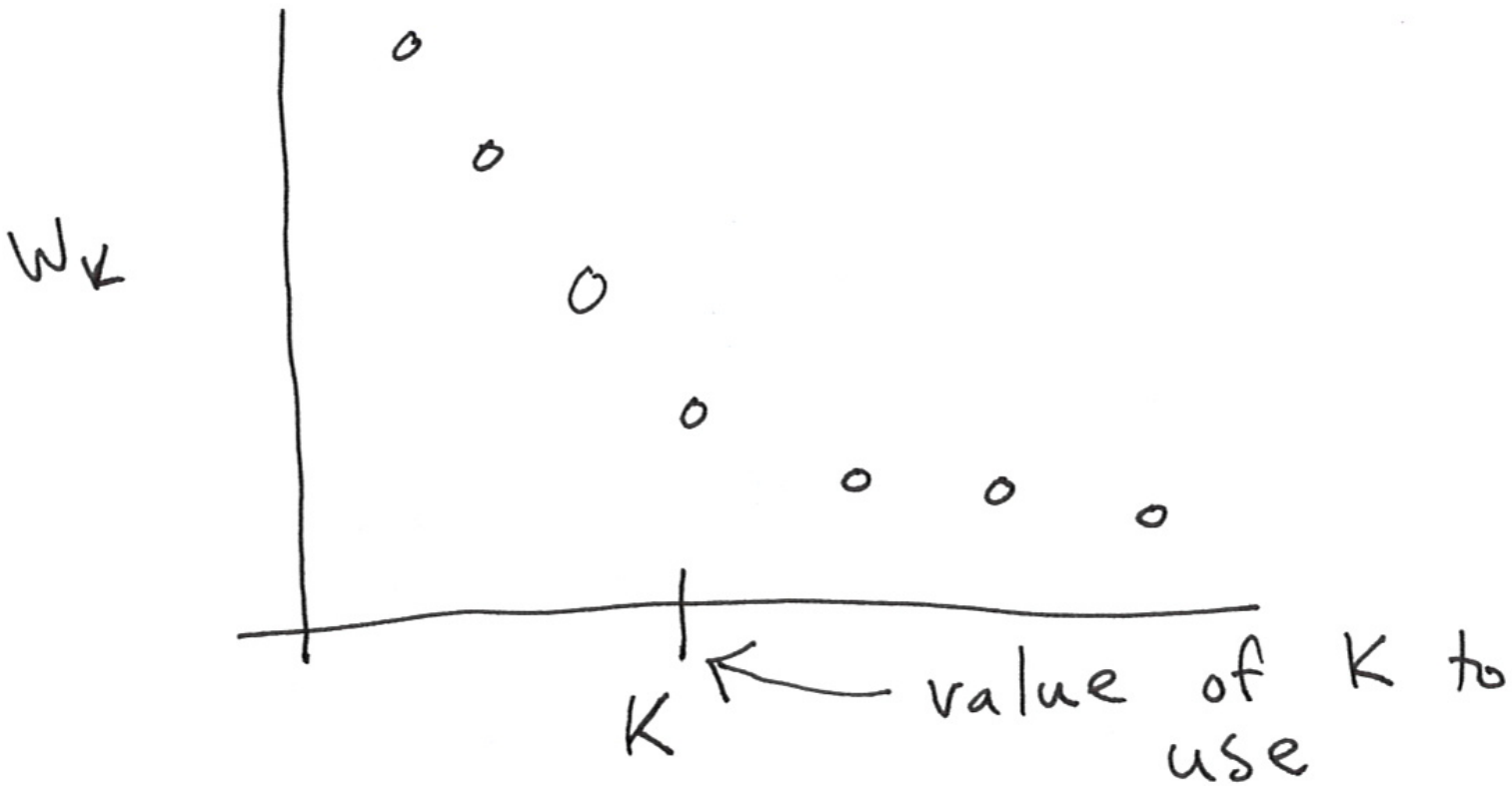
2. Look at the within-cluster variability W_K as a function of K

3.

Choosing the # Clusters

...so one can look for a “kink” in the plot of W_K versus K.

Choosing the # Clusters



Choosing the # Clusters

Option 3: Choose K based on the goals of the application (often the most practical choice)

Utilities example:

-
-

Fun Example

Loon Calls (project of a former 4740 student):

- Say you have recordings of the vocalizations of loons (a type of bird).
- A researcher has recorded the identity of the loon making each call. You develop a method to identify a loon based on its call.
- What type of method is yours?

  A. Regression
  B. Dimension Reduction
  C. Classification
  D. Clustering

Fun Example

- Now say you have recordings of loon calls where you do not know the identities of the loons making the calls. However, you know that a loon's call is distinctive and consistent
- You create a method that groups recorded calls that are similar, in the hopes that each group of calls corresponds to a single loon.
- What type of method is yours?

  A. Regression
  B. Dimension Reduction
  C. Classification
  D. Clustering