

Unsupervised Learning

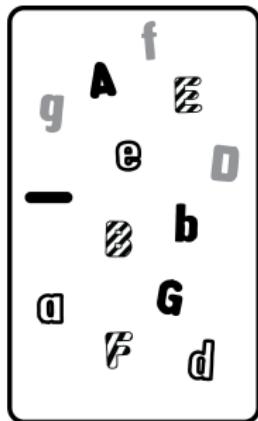
Bojan Božić

TU Dublin

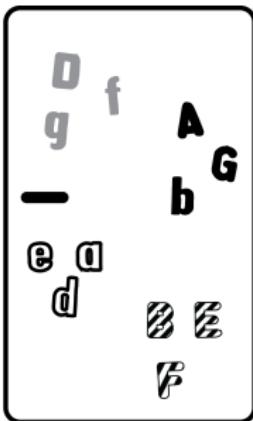
Summer 2021

- 1 Big Idea
- 2 Fundamentals
- 3 Standard Approach: The k -Means Clustering Algorithm
- 4 Extensions and Variations
- 5 Summary
- 6 Further Reading

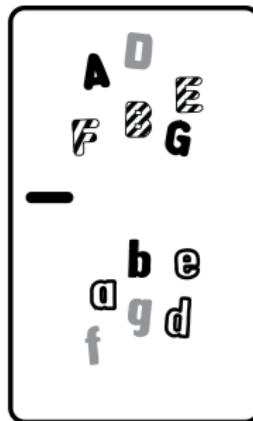
Big Idea



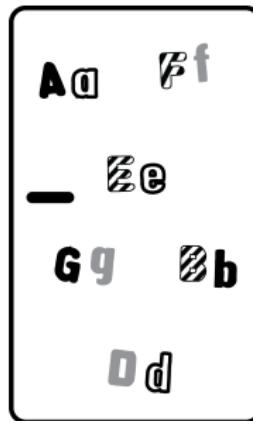
(a) The fridge



(b) Abigail



(c) Andrew



(d) Amalia

Figure 1: The three different arrangements of the magnetic letters made by the Murphy children on the Murphy family refrigerator.

Fundamentals

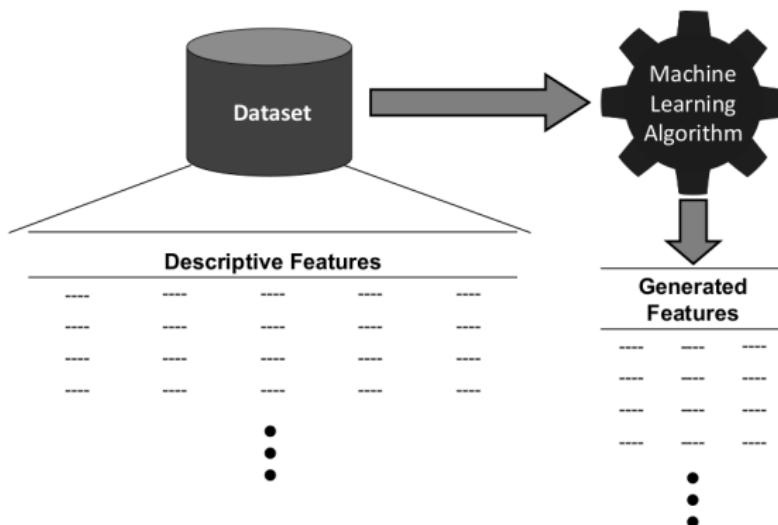


Figure 2: Unsupervised machine learning as a single-step process.

Standard Approach: The k -Means Clustering Algorithm



$$\sum_{i=1}^n \min_{\mathbf{c}_1, \dots, \mathbf{c}_k} Dist(\mathbf{d}_i, \mathbf{c}_j) \quad (1)$$

Pseudocode description of the ***k*-means clustering** algorithm.

Require: a dataset \mathcal{D} containing n training instances, $\mathbf{d}_1, \dots, \mathbf{d}_n$

Require: the number of clusters to find k

Require: a distance measure, $Dist$, to compare instances to cluster centroids

- 1: Select k random cluster centroids, \mathbf{c}_1 to \mathbf{c}_k , each defined by values for each descriptive feature, $\mathbf{c}_i = < \mathbf{c}_i[1], \dots, \mathbf{c}_i[m] >$
- 2: **repeat**
- 3: calculate the distance of each instance, \mathbf{d}_i , to each cluster centroid, \mathbf{c}_1 to \mathbf{c}_k , using $Dist$
- 4: assign each instance, \mathbf{d}_i , to belong to the cluster, \mathcal{C}_i , to whose cluster centroid, \mathbf{c}_i , it is closest
- 5: update each cluster centroid, \mathbf{c}_i , to the average of the descriptive feature values of the instances that belong to cluster \mathcal{C}_i
- 6: **until** no cluster reassessments are performed during an iteration

Table 1: A dataset of mobile phone customers described by their average monthly data (DATA USAGE) and call (CALL VOLUME) usage. Details of the first two iterations of the k -means clustering algorithm are also shown.

ID	DATA	CALL	Cluster Distances Iter. 1			Iter. 1 Cluster	Cluster Distances Iter. 2		
	USAGE	VOLUME	$Dist(\mathbf{d}_i, \mathbf{c}_1)$	$Dist(\mathbf{d}_i, \mathbf{c}_2)$	$Dist(\mathbf{d}_i, \mathbf{c}_3)$		$Dist(\mathbf{d}_i, \mathbf{c}_1)$	$Dist(\mathbf{d}_i, \mathbf{c}_2)$	$Dist(\mathbf{d}_i, \mathbf{c}_3)$
1	-0.9531	-0.3107	0.2341	0.9198	0.6193	\mathcal{C}_1	0.4498	1.9014	
2	-1.1670	-0.7060	0.5770	0.6108	0.9309	\mathcal{C}_1	0.87	2.0554	
3	-1.2329	-0.4188	0.3137	0.8945	0.6388	\mathcal{C}_1	0.7464	2.152	
4	1.0684	-0.4560	2.1972	2.06	2.438	\mathcal{C}_2	1.6857	0.3813	
5	-1.1104	0.1090	0.2415	1.3594	0.1973	\mathcal{C}_3	0.5669	2.1905	
6	-0.8431	0.1811	0.4084	1.405	0.4329	\mathcal{C}_1	0.3694	1.9842	
7	-0.3666	0.6905	1.1055	1.9728	1.0231	\mathcal{C}_3	0.7885	1.9406	
8	0.9285	-0.2168	2.0351	2.0378	2.2455	\mathcal{C}_1	1.5083	0.5759	
9	1.1175	-0.6028	2.2715	2.0566	2.529	\mathcal{C}_2	1.772	0.298	
10	0.8404	-1.0450	2.1486	1.693	2.4636	\mathcal{C}_2	1.7165	0.258	
11	-1.005	-0.0337	0.1404	1.2012	0.3692	\mathcal{C}_1	0.4339	2.0376	
12	0.2410	0.7360	1.6017	2.2398	1.6013	\mathcal{C}_3	1.1457	1.6581	
13	0.2021	0.4364	1.4253	1.9619	1.4925	\mathcal{C}_1	0.9259	1.4055	
14	0.2153	0.8360	1.6372	2.3159	1.6125	\mathcal{C}_3	1.2012	1.7602	
15	0.8770	-0.2459	1.985	1.9787	2.201	\mathcal{C}_2	1.4603	0.5454	
16	-0.0345	1.0502	1.595	2.4136	1.4929	\mathcal{C}_3	1.2433	2.0589	
17	0.8785	-1.3601	2.3325	1.727	2.6698	\mathcal{C}_2	1.9413	0.569	
18	0.9164	-0.8517	2.1454	1.7984	2.4383	\mathcal{C}_2	1.6815	0.0674	
19	-1.0423	0.1193	0.2593	1.3579	0.2525	\mathcal{C}_3	0.5065	2.133	
20	-0.7426	0.0119	0.3899	1.2399	0.5706	\mathcal{C}_1	0.1889	1.8164	
21	0.6259	-1.1834	2.0248	1.4696	2.3616	\mathcal{C}_2	1.6355	0.4709	
22	0.7684	-0.5844	1.927	1.7338	2.195	\mathcal{C}_2	1.4362	0.2382	
23	-0.2596	0.7450	1.2183	2.0535	1.1432	\mathcal{C}_3	0.8736	1.9167	
24	-0.3414	0.4215	0.9432	1.7202	0.9548	\mathcal{C}_1	0.5437	1.7259	

A Worked Example

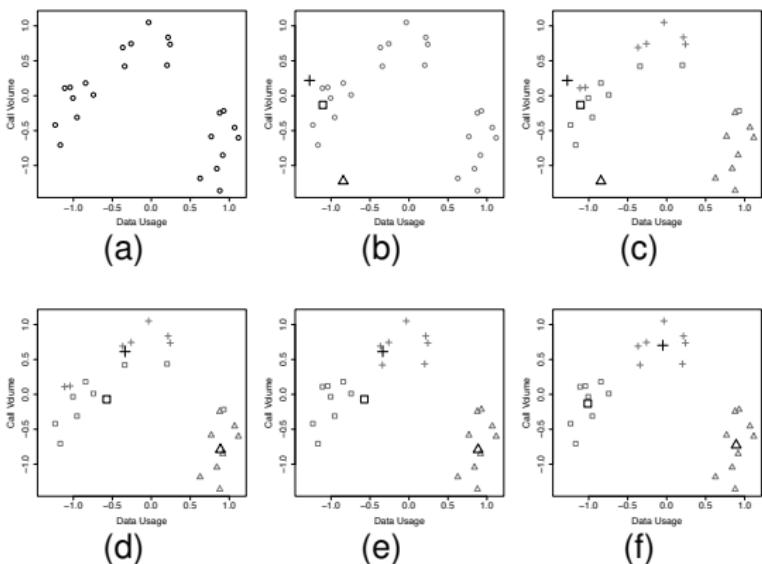


Figure 3: (a) A plot of the mobile phone customer dataset given in Table 1^[10]. (b)–(f) The progress of the k -means clustering algorithm, working on the simple customer segmentation dataset. The large symbols represent cluster centroids, and the smaller symbols represent cluster assignments.

A Worked Example

$$\mathbf{c}_1[\text{DATA USAGE}] = (-0.9531 + -1.167 + -1.2329 + -0.8431 + 0.9285 + -1.005 + 0.2021 + -0.7426 + -0.3414)/9$$

$$= -0.5727$$

$$\mathbf{c}_1[\text{CALL VOLUME}] = (-0.3107 + -0.706 + -0.4188 + 0.1811 + -0.2168 + -0.0337 + 0.4364 + 0.0119 + 0.4215)/9$$

$$= -0.0706$$

A Worked Example

$$\mathcal{C}_1 = \{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_5, \mathbf{d}_6, \mathbf{d}_{11}, \mathbf{d}_{19}, \mathbf{d}_{20}\}$$

$$\mathcal{C}_2 = \{\mathbf{d}_4, \mathbf{d}_8, \mathbf{d}_9, \mathbf{d}_{10}, \mathbf{d}_{15}, \mathbf{d}_{17}, \mathbf{d}_{18}, \mathbf{d}_{21}, \mathbf{d}_{22}\}$$

$$\mathcal{C}_3 = \{\mathbf{d}_7, \mathbf{d}_{12}, \mathbf{d}_{13}, \mathbf{d}_{14}, \mathbf{d}_{16}, \mathbf{d}_{23}, \mathbf{d}_{24}\}$$

Extensions and Variations

Choosing Initial Cluster Centroids

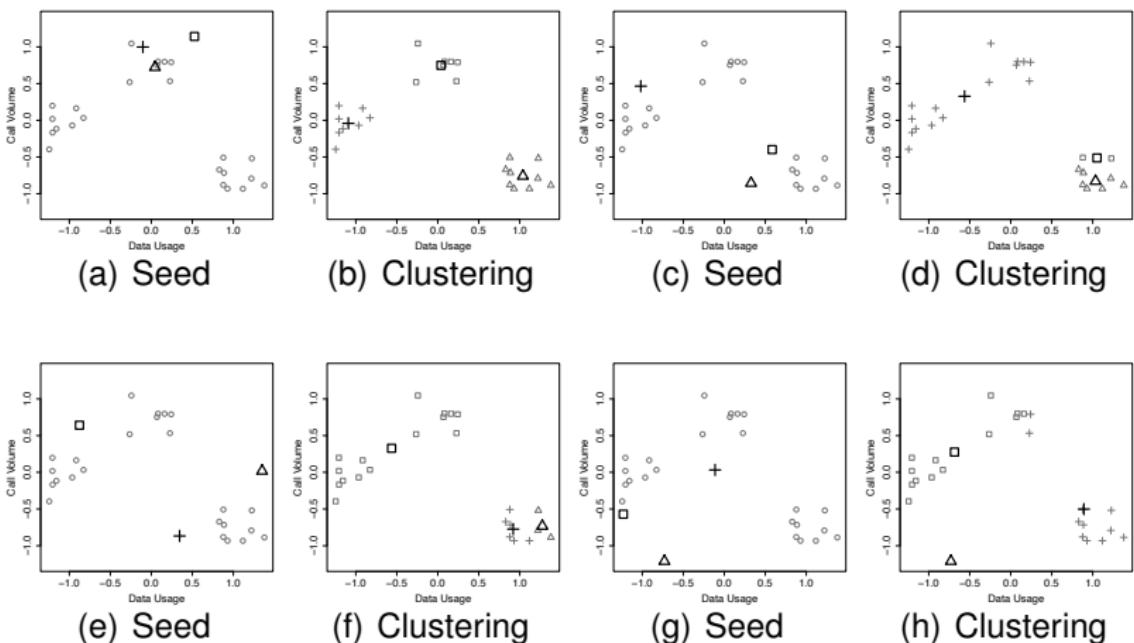


Figure 4: (a)–(h) Different clusterings (all with $k = 3$) that can be found for the mobile phone customer dataset given in Table 1^[10] when different initial cluster centroids are used.

Pseudocode description of the **k-means++** algorithm.

Require: a dataset \mathcal{D} containing n training instances, $\mathbf{d}_1, \dots, \mathbf{d}_n$

Require: k , the number of cluster centroids to find

Require: a distance measure $Dist$ to compare instances to cluster centroids

1: choose \mathbf{d}_i randomly (following a uniform distribution) from \mathcal{D} to be the position of the initial centroid, \mathbf{c}_1 , of the first cluster, \mathcal{C}_1

2: **for** cluster \mathcal{C}_j in \mathcal{C}_2 to \mathcal{C}_k **do**

3: for each instance, \mathbf{d}_i , in \mathcal{D} let $Dist(\mathbf{d}_i)$ be the distance between \mathbf{d}_i and its nearest cluster centroid

4: calculate a selection weight for each instance, \mathbf{d}_i , in \mathcal{D} as

$$\frac{Dist(\mathbf{d}_i)^2}{\sum_{p=1}^n Dist(\mathbf{d}_p)^2}$$

5: choose \mathbf{d}_i as the position of cluster centroid, \mathbf{c}_j , for cluster \mathcal{C}_j randomly following a distribution based on the selection weights

6: **end for**

7: proceed with k -means as normal using $\{\mathbf{c}_1, \dots, \mathbf{c}_k\}$ as the initial centroids.

Choosing Initial Cluster Centroids

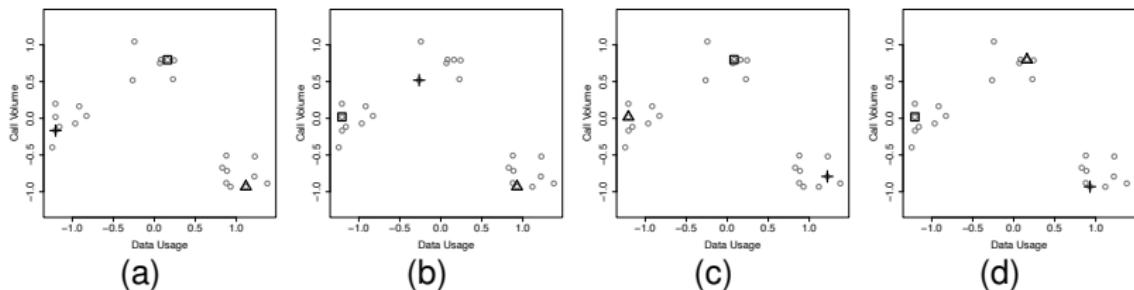


Figure 5: (a)–(d) Initial centroids chosen using the k-means++ approach (all with $k = 3$) for the mobile phone customer dataset given in Table 1^[10].

Evaluating Clustering

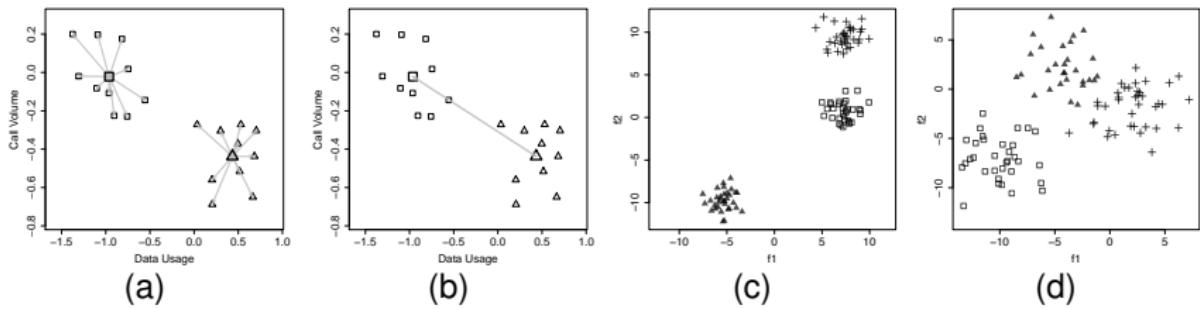


Figure 6: (a) Intra-cluster distance; (b) inter-cluster distance; (c) a *good* clustering; and (d) a *bad* clustering.

Evaluating Clustering

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2)$$

Evaluating Clustering

	d_2	d_3	d_5	d_6	d_{11}	d_{19}	d_{20}		
d_1	[0.45	0.30	0.45	0.50	0.28	0.44	0.39]		
	d_4	d_8	d_9	d_{10}	d_{15}	d_{17}	d_{18}	d_{21}	d_{22}
d_1	[2.03	1.88	2.09	1.94	1.83	2.11	1.95	1.80	1.74]
	d_7	d_{12}	d_{13}	d_{14}	d_{16}	d_{23}	d_{24}		
d_1	[1.16	1.59	1.38	1.64	1.64	1.26	0.95]		

Evaluating Clustering

$$\frac{1.3743 - 0.401}{\max(0.401, 1.374)} = 0.7081$$

Pseudocode description of the algorithm for calculating the **silhouette** for internal cluster evaluation.

Require: a dataset \mathcal{D} containing n training instances, $\mathbf{d}_1, \dots, \mathbf{d}_n$

Require: a clustering \mathcal{C} of dataset \mathcal{D} into k clusters, $\mathcal{C}_1, \dots, \mathcal{C}_k$

Require: a distance measure, $Dist$, to compare distances between instances

- 1: **for** each instance \mathbf{d}_i in \mathcal{D} **do**
- 2: let $a(i)$ be the average distance between instance \mathbf{d}_i and all of the other instances within the cluster to which \mathbf{d}_i belongs, \mathcal{C}_j (*average intra-cluster distance*)
- 3: calculate the average distance between instance \mathbf{d}_i and the members of each of the other clusters $\mathcal{C} \setminus \mathcal{C}_j$
- 4: let $b(i)$ be the lowest average distance between instance \mathbf{d}_i and any other cluster (*average inter-cluster distance*)
- 5: calculate the silhouette index for \mathbf{d}_i as
$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (3)$$
- 6: **end for**
- 7: calculate final silhouette for the clustering as $s = \frac{1}{n} \sum_{i=1}^n s(i)$

Evaluating Clustering

Table 2: Calculating the silhouette for the final clustering of the mobile phone customer dataset (Table 1^[10]) found using the k -means algorithm (with $k = 3$). The overall silhouette index value is 0.66.

ID	Cluster	Nearest Cluster			Nearest Cluster						
		$a(i)$	$b(i)$	$s(i)$	$a(i)$	$b(i)$	$s(i)$				
1	\mathcal{C}_1	\mathcal{C}_3	0.401	1.374	0.708	13	\mathcal{C}_3	\mathcal{C}_1	0.5136	1.3592	0.6221
2	\mathcal{C}_1	\mathcal{C}_3	0.695	1.811	0.616	14	\mathcal{C}_3	\mathcal{C}_1	0.4349	1.5738	0.7236
3	\mathcal{C}_1	\mathcal{C}_3	0.503	1.644	0.694	15	\mathcal{C}_2	\mathcal{C}_3	0.5776	1.3480	0.5715
4	\mathcal{C}_2	\mathcal{C}_3	0.484	1.628	0.703	16	\mathcal{C}_3	\mathcal{C}_1	0.4955	1.5409	0.6784
5	\mathcal{C}_1	\mathcal{C}_3	0.387	1.232	0.686	17	\mathcal{C}_2	\mathcal{C}_1	0.7369	2.2757	0.6762
6	\mathcal{C}_1	\mathcal{C}_3	0.445	0.970	0.541	18	\mathcal{C}_2	\mathcal{C}_3	0.4312	1.8473	0.7666
7	\mathcal{C}_3	\mathcal{C}_1	0.452	1.056	0.572	19	\mathcal{C}_1	\mathcal{C}_3	0.3711	1.1682	0.6823
8	\mathcal{C}_2	\mathcal{C}_3	0.599	1.364	0.561	20	\mathcal{C}_1	\mathcal{C}_3	0.4334	1.0006	0.5669
9	\mathcal{C}_2	\mathcal{C}_3	0.470	1.768	0.734	21	\mathcal{C}_2	\mathcal{C}_1	0.6520	1.9710	0.6692
10	\mathcal{C}_2	\mathcal{C}_3	0.504	1.978	0.745	22	\mathcal{C}_2	\mathcal{C}_3	0.4504	1.5457	0.7086
11	\mathcal{C}_1	\mathcal{C}_3	0.327	1.223	0.732	23	\mathcal{C}_3	\mathcal{C}_1	0.3954	1.1654	0.6607
12	\mathcal{C}_3	\mathcal{C}_1	0.433	1.537	0.719	24	\mathcal{C}_3	\mathcal{C}_1	0.5339	0.8880	0.3988

Evaluating Clustering

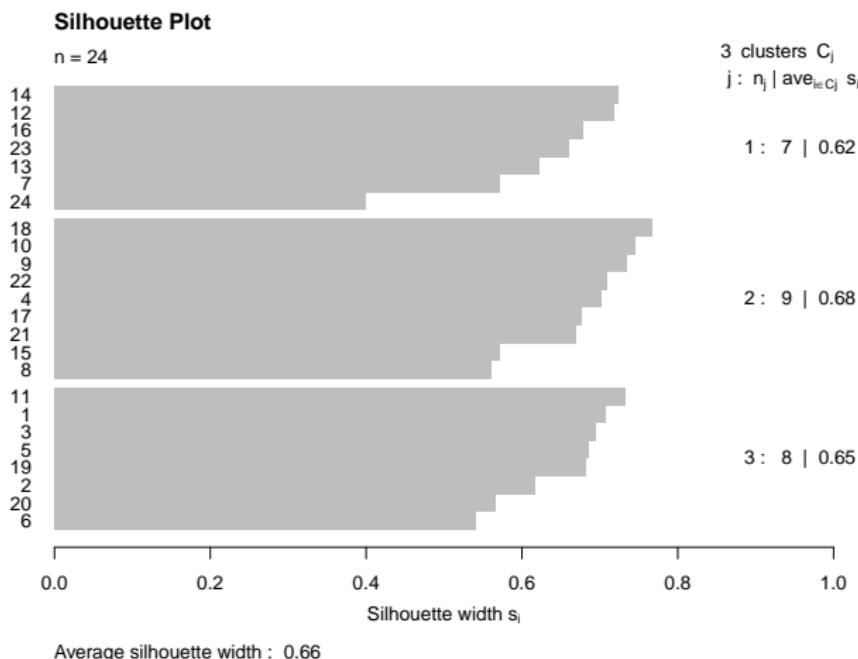


Figure 7: The silhouette plot for the final clustering of the mobile phone customer dataset (Table 1^[10]) found using the k -means algorithm (with $k = 3$).

Choosing the Number of Clusters

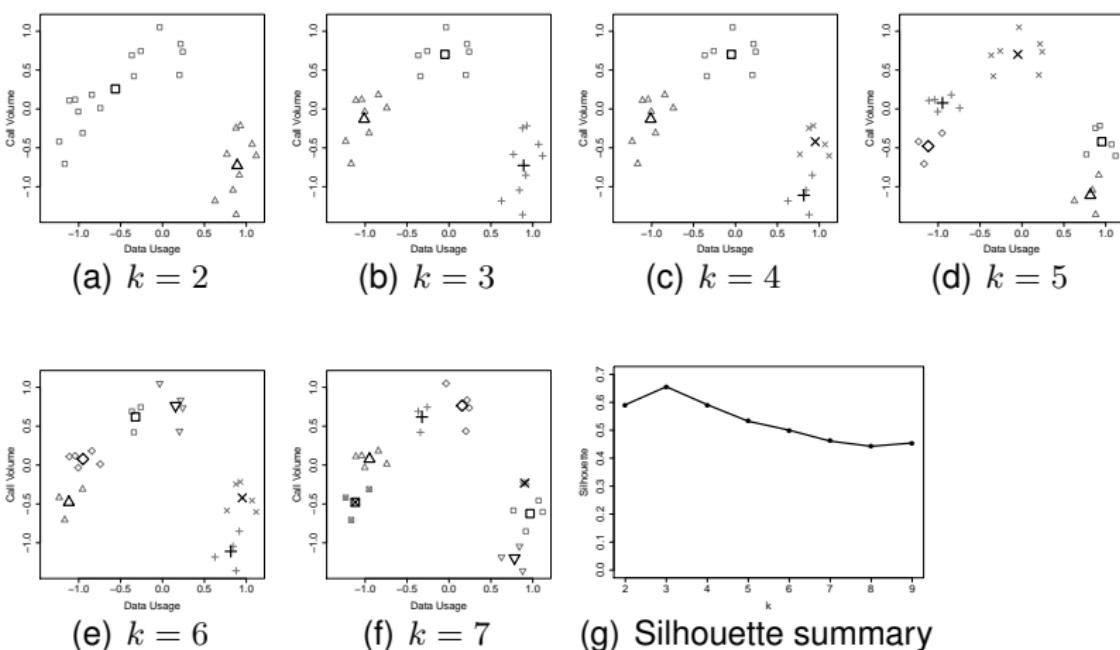


Figure 8: (a)–(f) Different clusterings found for the mobile phone customer dataset in Table 1^[10] for values of k in $(2, 9)$. (g) shows the silhouette for each clustering.

Understanding Clustering Results

Table 3: Summary statistics for the three clusters found in the mobile phone customer dataset in Table 1^[10] using k -means clustering ($k = 3$). Note, that the % missing and cardinality columns usually used are omitted here for legibility as these data quality issues will not arise in this simple example. They could be included when this approach is used on *real* datasets.

Feature	Cluster	Count	1 st				3 rd			Std. Dev.
			Min.	Qrt.	Mean	Median	Qrt.	Max		
DATA USAGE	\mathcal{C}_1	8	-1.2329	-1.1246	-1.0121	-1.0237	-0.9256	-0.7426	0.1639	
	\mathcal{C}_2	9	0.6259	0.8404	0.8912	0.8785	0.9285	1.1175	0.1471	
	\mathcal{C}_3	7	-0.3666	-0.3005	-0.0491	-0.0345	0.2087	0.241	0.2732	
CALL VOLUME	\mathcal{C}_1	8	-0.7060	-0.3377	-0.1310	-0.0109	0.1116	0.1811	0.3147	
	\mathcal{C}_2	9	-1.3601	-1.0450	-0.7273	-0.6028	-0.4560	-0.2168	0.4072	
	\mathcal{C}_3	7	0.4215	0.5635	0.7022	0.7360	0.7905	1.0502	0.2204	

Understanding Clustering Results

Table 4: Information gain for each descriptive feature as a predictor of membership of each cluster based on the clustering of the mobile phone customer dataset in Table 1^[10] found using k -means clustering ($k = 3$).

Feature	C_1	C_2		C_3	
	Info. Gain	Feature	Info. Gain	Feature	Info. Gain
DATA.USAGE	0.9183	DATA.USAGE	0.9544	CALL.VOLUME	0.8709
CALL.VOLUME	0.2117	CALL.VOLUME	0.5488	DATA.USAGE	0.2479

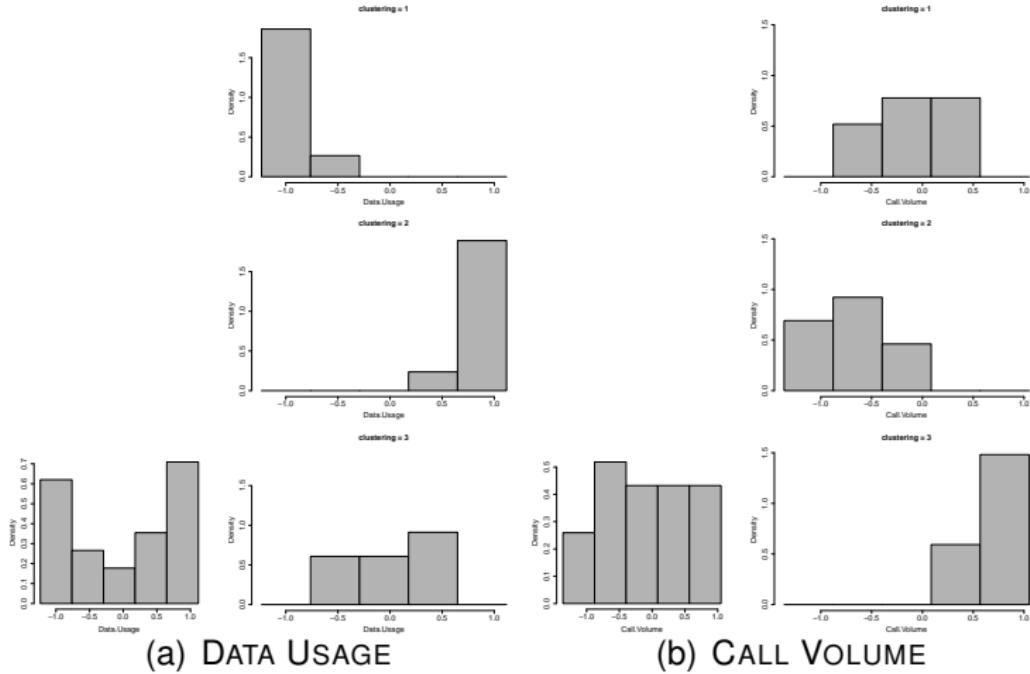


Figure 9: (a)–(b) Visualizations of the distributions of the descriptive features in the mobile phone customer dataset in Table 1^[10] across the complete dataset, and divided by the clustering found using k -means clustering ($k = 3$).

Agglomerative Hierarchical Clustering

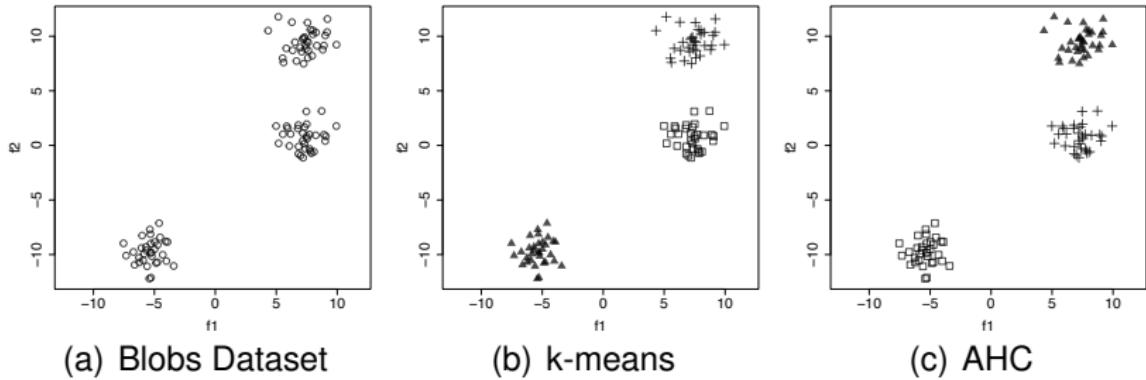
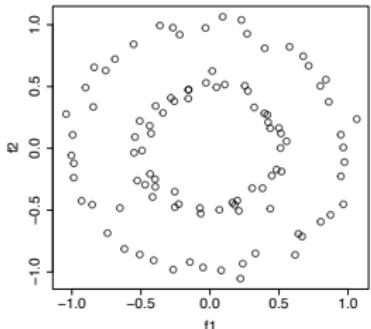
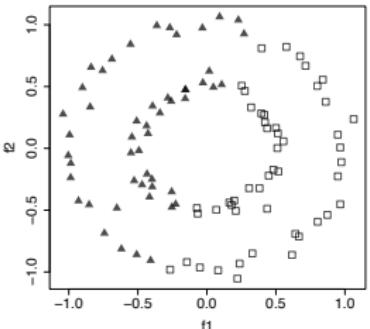


Figure 10: (a)–(i) A plot of the *blobs*, *circles*, and *half-moons* datasets and the clusterings achieved by the k -means clustering and agglomerative hierarchical clustering algorithms (where k is set to 3, 2, and 2, respectively).

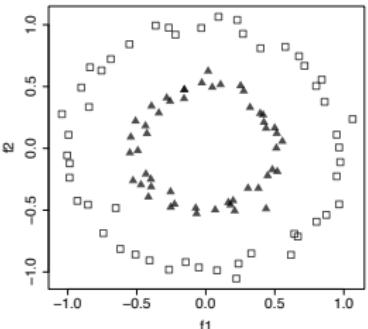
Agglomerative Hierarchical Clustering



(d) Circles Dataset



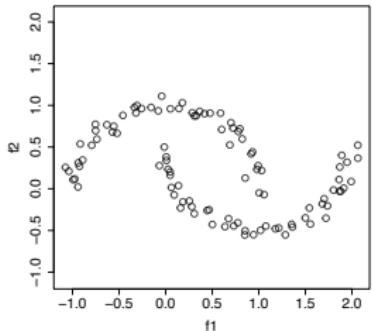
(e) k-means



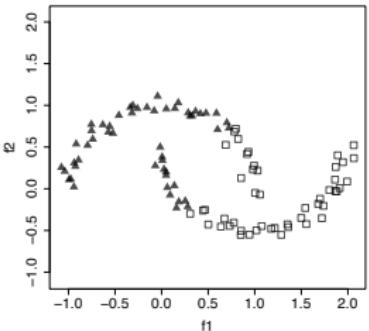
(f) AHC

Figure 11: (a)–(i) A plot of the *blobs*, *circles*, and *half-moons* datasets and the clusterings achieved by the k -means clustering and agglomerative hierarchical clustering algorithms (where k is set to 3, 2, and 2, respectively).

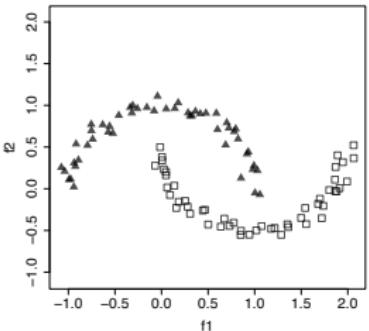
Agglomerative Hierarchical Clustering



(g) Half-moons Dataset



(h) k-means



(i) AHC

Figure 12: (a)–(i) A plot of the *blobs*, *circles*, and *half-moons* datasets and the clusterings achieved by the k -means clustering and agglomerative hierarchical clustering algorithms (where k is set to 3, 2, and 2, respectively).

Pseudocode description of the **agglomerative hierarchical clustering** algorithm.

Require: a dataset \mathcal{D} containing n training instances, $\mathbf{d}_1, \dots, \mathbf{d}_n$

Require: a distance measure, $Dist$, to compare distances between instances

Require: a linkage method, \mathcal{L} , to compare distances between clusters

- 1: initialize the hierarchy level, $h = 1$
- 2: divide \mathcal{D} into a set of n disjoint clusters, $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_n\}$, with one instance in each cluster
- 3: **repeat**
- 4: using distance measure $Dist$ and linkage method \mathcal{L} , find the nearest pair of clusters, \mathcal{C}_i and \mathcal{C}_j , in the current clustering
- 5: merge \mathcal{C}_i and \mathcal{C}_j to form a new cluster \mathcal{C}_{n+h}
- 6: remove the old clusters from the clustering: $\mathcal{C} \leftarrow \mathcal{C} \setminus \{\mathcal{C}_i, \mathcal{C}_j\}$
- 7: add the new cluster to the clustering: $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}_{n+h}$
- 8: $h \leftarrow h + 1$
- 9: **until** all the instances join into a single cluster

Agglomerative Hierarchical Clustering

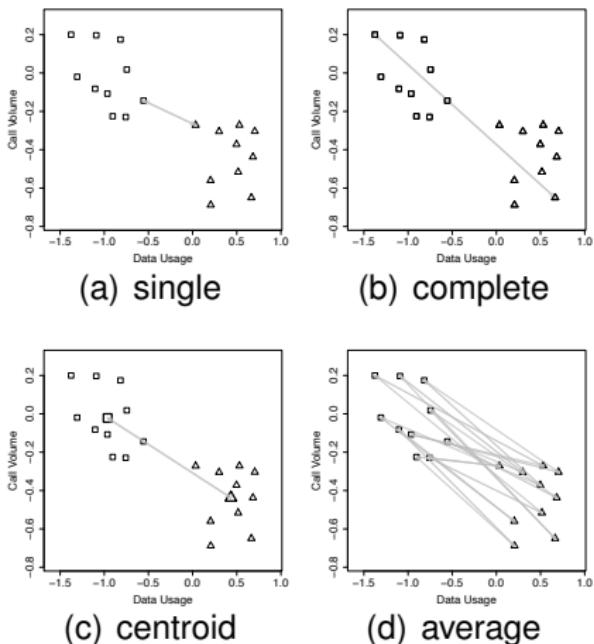


Figure 13: (a)–(d) Different linkage methods that can be used to compare the distances between clusters in agglomerative hierarchical clustering. (Arrows for only some indicative distances are shown in the average linkage diagram (d).)

Table 5: Distance matrices that detail the first three iterations of the AHC algorithm applied to the reduced version of the mobile phone customer dataset in Table 1^[10].

(a) A distance matrix for the instances in the dataset.

	d_4	d_{15}	d_8	d_{11}	d_5	d_{19}	d_{24}	d_7	d_{23}
d_4	0.00								
d_{15}	0.28	0.00							
d_8	0.28	0.06	0.00						
d_{11}	2.12	1.89	1.94	0.00					
d_5	2.25	2.02	2.06	0.18	0.00				
d_{19}	2.19	1.95	2.00	0.16	0.07	0.00			
d_{24}	1.66	1.39	1.42	0.81	0.83	0.76	0.00		
d_7	1.84	1.56	1.58	0.96	0.94	0.89	0.27	0.00	
d_{23}	1.79	1.51	1.53	1.08	1.06	1.00	0.33	0.12	0.00

(b) The distance matrix after one iteration of AHC.

	d_4	c_{10}	d_{11}	d_5	d_{19}	d_{24}	d_7	d_{23}
d_4	0.00							
c_{10}	0.28		0.00					
d_{11}	2.12		1.89	0.00				
d_5	2.25		2.02	0.18	0.00			
d_{19}	2.19		1.95	0.16	0.07	0.00		
d_{24}	1.66		1.39	0.81	0.83	0.76	0.00	
d_7	1.84		1.56	0.96	0.94	0.89	0.27	0.00
d_{23}	1.79		1.51	1.08	1.06	1.00	0.33	0.12



Table 6: Distance matrices that detail the first three iterations of the AHC algorithm applied to the reduced version of the mobile phone customer dataset in Table 1^[10].

(c) The distance matrix after two iterations of AHC.

	d_4	C_{10}	d_{11}	C_{11}	d_{24}	C_{12}
d_4	0.00					
C_{10}	0.28		0.00			
d_{11}	2.12		1.89 0.00			
C_{11}	2.19		1.95 0.16	0.00		
d_{24}	1.66		1.39 0.81	0.76 0.00		
C_{12}	1.79		1.51 0.97	0.89 0.27	0.00	

(d) The distance matrix after three iterations of AHC.

	d_4	C_{13}	C_{11}	d_{24}	C_{12}
d_4	0.00				
C_{13}	0.28		0.00		
C_{11}	2.19		0.16	0.00	
d_{24}	1.66		0.81	0.76 0.00	
C_{12}	1.79		0.97	0.89 0.27	0.00

Agglomerative Hierarchical Clustering

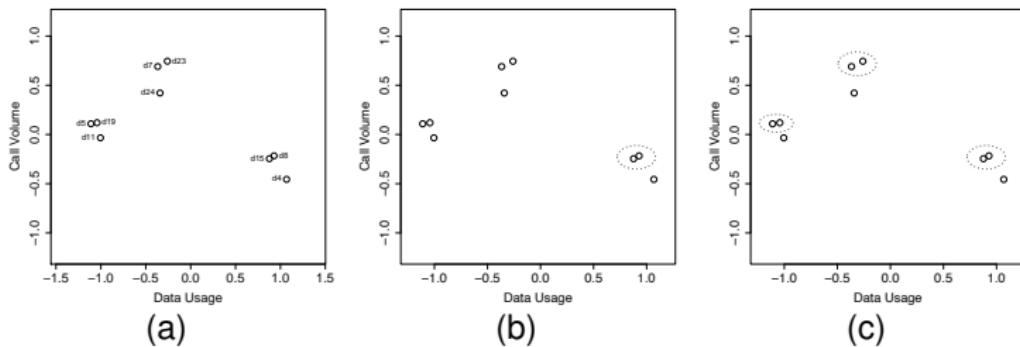


Figure 14: (a) A plot of a reduced version of the mobile phone customer dataset given in Table 1^[10]. (b) At the first iteration of the AHC algorithm the first pair of instances is combined into a cluster, \mathcal{C}_{10} . (c) After three iterations of the AHC algorithm, three pairs of instances have been combined into clusters, \mathcal{C}_{10} , \mathcal{C}_{11} , and \mathcal{C}_{12} . (d) At the fourth iteration of AHC, the first hierarchical cluster combination is created when a single instance, d_{11} is combined with the cluster \mathcal{C}_{10} to create a new cluster, \mathcal{C}_{13} .

Agglomerative Hierarchical Clustering

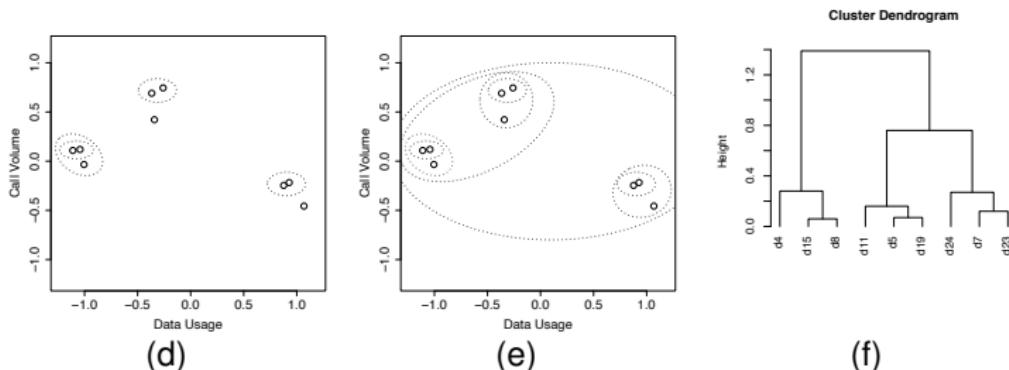


Figure 15: (a) A plot of a reduced version of the mobile phone customer dataset given in Table 1^[10]. (b) At the first iteration of the AHC algorithm the first pair of instances is combined into a cluster, \mathcal{C}_{10} . (c) After three iterations of the AHC algorithm, three pairs of instances have been combined into clusters, \mathcal{C}_{10} , \mathcal{C}_{11} , and \mathcal{C}_{12} . (d) At the fourth iteration of AHC, the first hierarchical cluster combination is created when a single instance, d_{11} is combined with the cluster \mathcal{C}_{10} to create a new cluster, \mathcal{C}_{13} .

Agglomerative Hierarchical Clustering

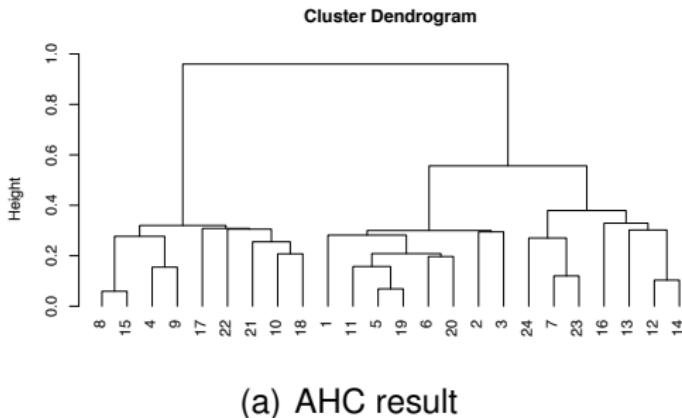


Figure 16: (a) A plot of the hierarchical grouping of the instances in the mobile phone customer dataset from Table 1^[10] found by the AHC algorithm (using Euclidean distance and single linkage). (b) The clustering returned when the tree is cut at $k = 3$. (c) The clustering returned when the tree is cut at $k = 6$.

Agglomerative Hierarchical Clustering

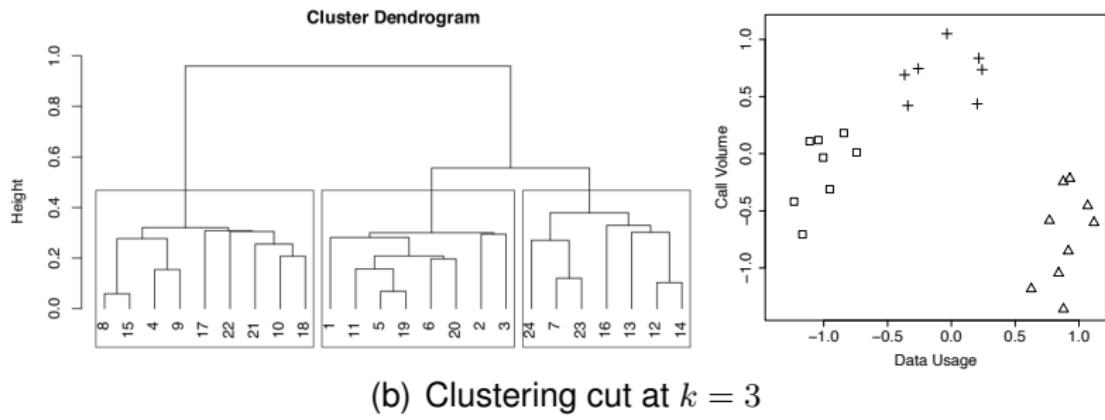


Figure 17: (a) A plot of the hierarchical grouping of the instances in the mobile phone customer dataset from Table 1^[10] found by the AHC algorithm (using Euclidean distance and single linkage). (b) The clustering returned when the tree is cut at $k = 3$. (c) The clustering returned when the tree is cut at $k = 6$.

Agglomerative Hierarchical Clustering

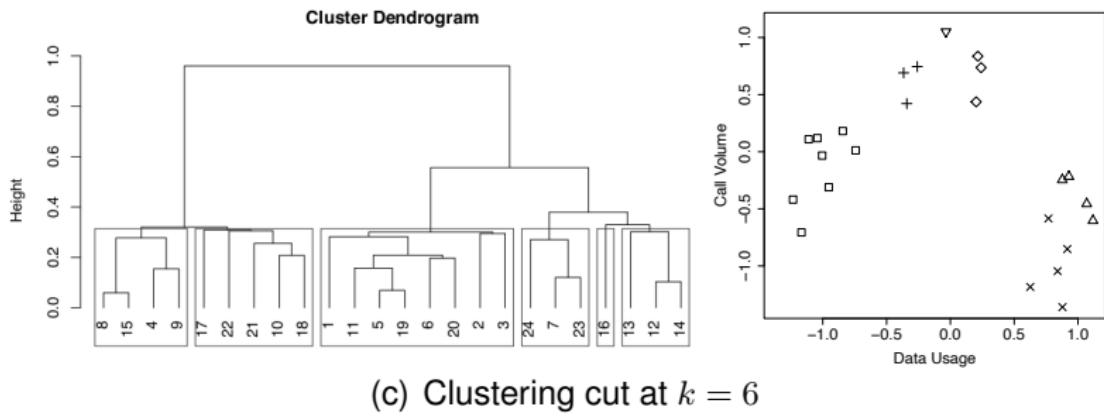
(c) Clustering cut at $k = 6$

Figure 18: (a) A plot of the hierarchical grouping of the instances in the mobile phone customer dataset from Table 1^[10] found by the AHC algorithm (using Euclidean distance and single linkage). (b) The clustering returned when the tree is cut at $k = 3$. (c) The clustering returned when the tree is cut at $k = 6$.

Representation Learning with Auto-Encoders

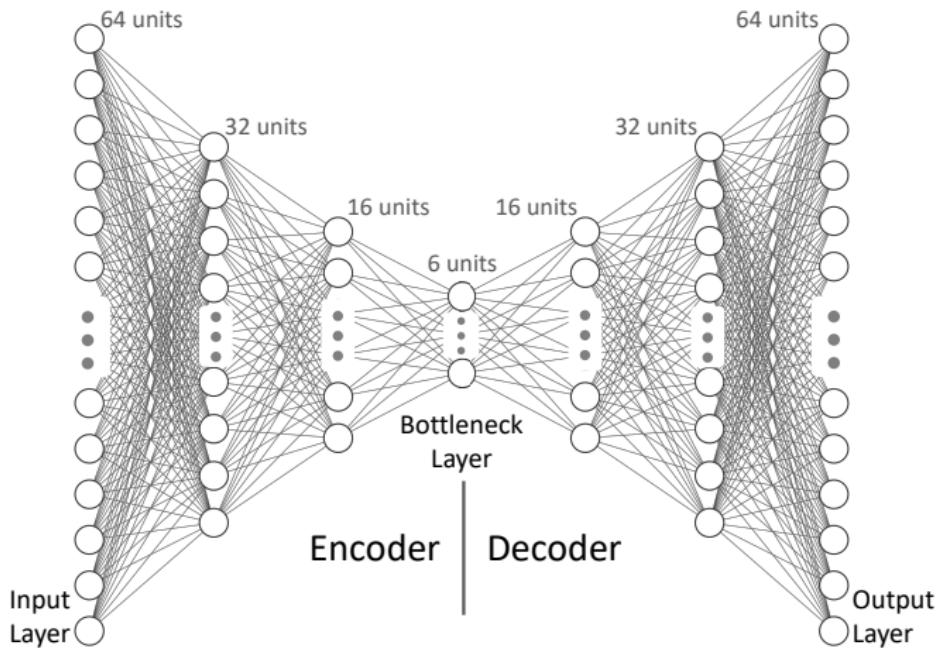


Figure 19: The architecture of an auto-encoder network made up of an encoder and a decoder connected by a bottleneck layer.



(a)



(b)



(c)



(d)

Figure 20: (a) A selection of images from the handwritten digits dataset; (b) image reconstructions generated by the auto-encoder network before training; (c) image reconstructions generated by the auto-encoder network after minimal training (10 epochs); and (d) image reconstructions generated by the auto-encoder network after complete training (1,000 epochs).

Training	Image	Pixel Values								Error
		[0.00 0.19 0.94 1.00 0.88 0.06 0.00 0.00]								
		[0.00 0.12 0.75 0.81 1.00 0.25 0.00 0.00]								
		[0.00 0.00 0.00 0.38 1.00 0.19 0.00 0.00]								
		[0.00 0.00 0.06 0.94 0.62 0.00 0.00 0.00]								
Original		[0.00 0.00 0.38 1.00 0.25 0.00 0.00 0.00]								
		[0.00 0.12 0.94 0.62 0.00 0.00 0.00 0.00]								
		[0.00 0.25 1.00 0.69 0.50 0.69 0.19 0.00]								
		[0.00 0.19 1.00 1.00 0.75 0.19 0.00 0.00]								
		[0.51 0.48 0.49 0.49 0.51 0.50 0.49 0.50]								
0 Epochs		[0.51 0.50 0.49 0.52 0.51 0.51 0.51 0.49]								0.1876
		[0.51 0.52 0.50 0.51 0.50 0.49 0.52 0.50]								
		[0.50 0.51 0.51 0.51 0.50 0.50 0.49 0.50]								
		[0.50 0.47 0.50 0.52 0.51 0.50 0.52 0.50]								
		[0.51 0.51 0.50 0.49 0.53 0.50 0.51 0.49]								
		[0.51 0.52 0.49 0.51 0.51 0.50 0.51 0.50]								
		[0.50 0.50 0.48 0.51 0.50 0.51 0.51 0.51]								

Figure 21: An image of the digit 2 and reconstructions of this image by the auto-encoder after various amounts of network training. The pixel values of the reconstructed images are shown alongside the images, as is the reconstruction error calculated by comparing these to the pixel values of the original image.

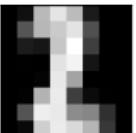
Training	Image	Pixel Values							Error
10 Epochs		0.00	0.00	0.42	0.83	0.76	0.33	0.03	0.00
		0.01	0.12	0.80	0.76	0.74	0.62	0.06	0.00
		0.00	0.15	0.60	0.35	0.54	0.58	0.07	0.00
		0.00	0.09	0.49	0.57	0.68	0.46	0.11	0.00
		0.00	0.08	0.31	0.57	0.68	0.48	0.13	0.01
		0.00	0.05	0.31	0.37	0.41	0.51	0.19	0.00
		0.00	0.02	0.49	0.59	0.59	0.63	0.21	0.01
		0.00	0.01	0.45	0.85	0.74	0.39	0.09	0.01
1,000 Epochs		0.00	0.06	0.71	0.87	0.71	0.10	0.00	0.00
		0.00	0.32	0.70	0.77	0.86	0.30	0.00	0.00
		0.00	0.11	0.09	0.75	0.97	0.24	0.00	0.00
		0.00	0.00	0.00	0.86	0.93	0.08	0.00	0.00
		0.00	0.00	0.02	0.88	0.62	0.00	0.00	0.00
		0.00	0.01	0.68	0.89	0.24	0.04	0.01	0.00
		0.00	0.32	0.91	0.89	0.53	0.51	0.19	0.00
		0.00	0.03	0.78	0.89	0.83	0.69	0.21	0.00

Figure 22: An image of the digit 2 and reconstructions of this image by the auto-encoder after various amounts of network training. The pixel values of the reconstructed images are shown alongside the images, as is the reconstruction error calculated by comparing these to the pixel values of the original image.

Representation Learning with Auto-Encoders

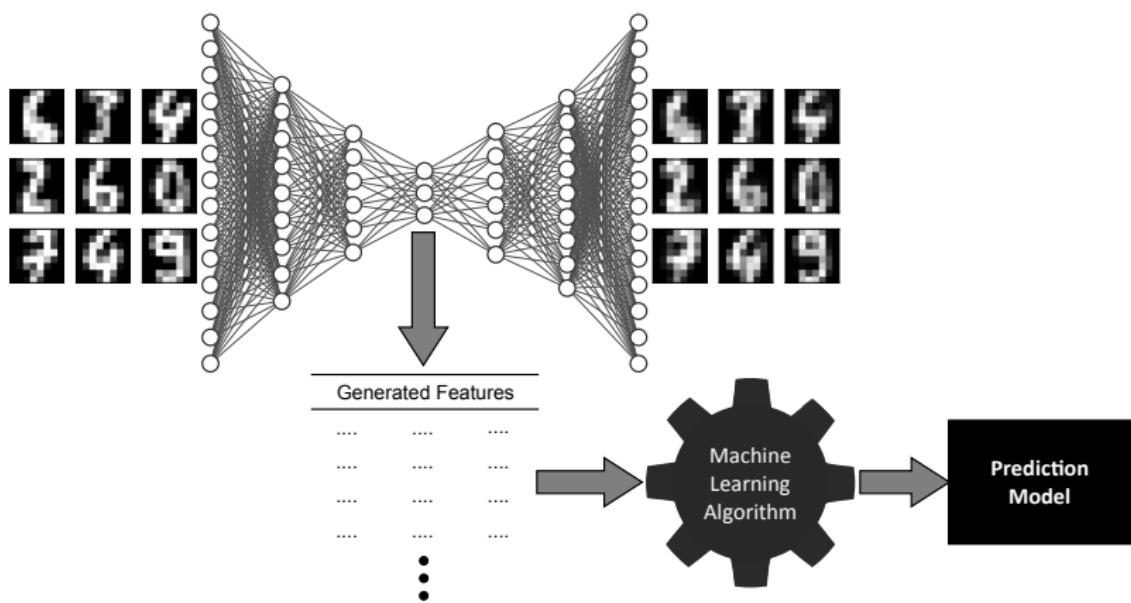


Figure 23: The process of using an unsupervised auto-encoder network to generate a feature representation used to train a supervised model.

Summary

- Unsupervised machine learning techniques are used in the absence of a target feature and model the underlying structure within the descriptive features in a dataset. W
- We can think of the output of most unsupervised machine learning models as new generated features that can be appended to the original dataset to **augment** or **enrich** it.
- There are two key main use cases for unsupervised learning: clustering and representation learning.
- Two clustering techniques were presented in detail:
k-means clustering and ***agglomerative hierarchical clustering*** (AHC).

- **Neural network** models are especially effective for representation learning, and the chapter presented an example of using an **auto-encoder** to learn a feature representation that could be used by a supervised machine learning model.
- Applications of unsupervised learning are widespread, including customer segmentations, anomaly detection, and analyzing people's movement patterns.
- Designing solutions based on unsupervised machine learning techniques can be quite creative.
- Finally, unsupervised learning is a fascinating research area and has many significant open research challenges.

Further Reading

- For more detail on unsupervised machine learning algorithms (Friedman et al., 2001) has a fairly comprehensive unsupervised learning section.
- For good treatments of unsupervised learning applications see (Berry and Linoff, 2004; Han et al., 2011).
- (Guo et al., 2016) provides a readable, coherent overview of different types of autoencoder models.

- 1 Big Idea
- 2 Fundamentals
- 3 Standard Approach: The k -Means Clustering Algorithm
- 4 Extensions and Variations
- 5 Summary
- 6 Further Reading

- Berry, Michael J. A., and Gordon S. Linoff. 2004. *Data mining techniques: for marketing, sales, and customer relationship management*. Wiley.
- Friedman, J., T. Hastie, and R. Tibshirani. 2001. *The elements of statistical learning*, Vol. 1. Springer.
- Guo, Yanming, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S. Lew. 2016. Deep learning for visual understanding: A review. *Neurocomputing* 187: 27–48.
- Han, Jiawei, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.