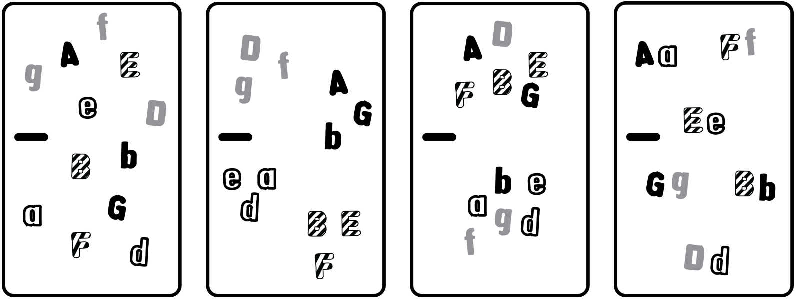
Beyond Prediction: Unsupervised Learning

John D. Kelleher and Brian Mac Namee and Aoife D'Arcy

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

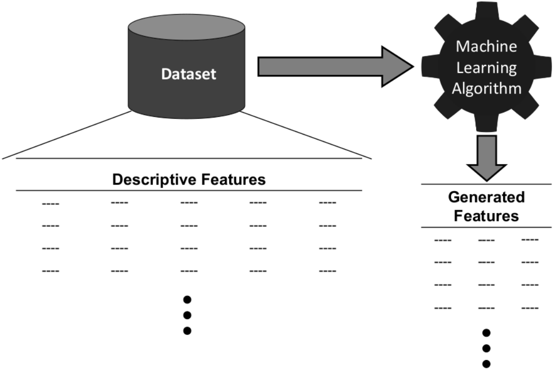


Figure 2: Unsupervised machine learning as a single-step process.

Standard Approach: The k-Means Clustering Algorithm

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

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

Require: a dataset $\mathcal{D}$ containing n training instances, $\mathbf{d}_1, \ldots, \mathbf{d}_n$
Require: the number of clusters to find k
Require: a distance measure, $Dist$, to compare instances to cluster centroids

```
Select k random cluster centroids, c_1 to c_k, each defined by values for each descriptive feature, c_i =< c_i[1], ..., c_i[m] >
repeat
    calculate the distance of each instance, d_i, to each cluster centroid, c_1 to c_k, using Dist
    assign each instance, d_i, to belong to the cluster, C_i, to whose cluster centroid, c_i, it is closest
    update each cluster centroid, c_i, to the average of the descriptive feature values of the instances that belong to cluster C_i
until no cluster reassignments 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 USAGE	CALL VOLUME	Cluster Distances Iter. 1			Iter. 1	Cluster Distances Iter.	
			$Dist(\mathbf{d}_i, \mathbf{c}_1)$	$Dist(\mathbf{d}_i, \mathbf{c}_2)$	$Dist(\mathbf{d}_i, \mathbf{c}_3)$	Cluster	$Dist(\mathbf{d}_i, \mathbf{c}_1)$	$Dist(\mathbf{d}_i, \mathbf{c}_2)$
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

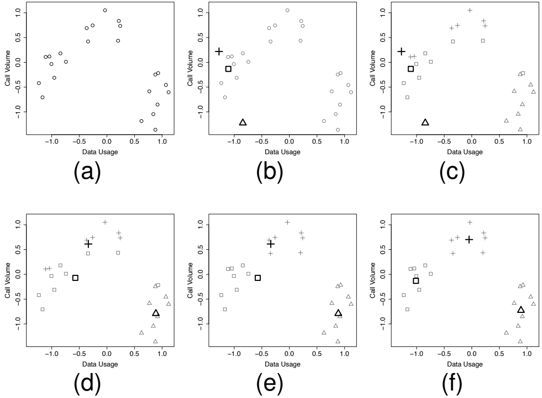


Figure 3: (a) A plot of the mobile phone customer dataset given in Table 1[10]. (b)–(f) The progress of the k-means clustering algorithm, working on the simple customer segmentation dataset. The large symbols represent cluster centroids, and the smaller symbols represent cluster assignments.

A Worked Example

$$
\begin{aligned}
\mathbf{c}_1[\textsc{Data Usage}] &= (-0.9531 + -1.167 + -1.2329 + -0.8431 + 0.9285 \\
&\qquad + -1.005 + 0.2021 + -0.7426 + -0.3414)/9 \\
&= -0.5727 \\
\mathbf{c}_1[\textsc{Call Volume}] &= (-0.3107 + -0.706 + -0.4188 + 0.1811 + -0.2168 \\
&\qquad + -0.0337 + 0.4364 + 0.0119 + 0.4215)/9 \\
&= -0.0706
\end{aligned}
$$

$$\begin{aligned}
\mathcal{C}_1 &= \{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_5, \mathbf{d}_6, \mathbf{d}_{11}, \mathbf{d}_{19}, \mathbf{d}_{20}\} \\
\mathcal{C}_2 &= \{\mathbf{d}_4, \mathbf{d}_8, \mathbf{d}_9, \mathbf{d}_{10}, \mathbf{d}_{15}, \mathbf{d}_{17}, \mathbf{d}_{18}, \mathbf{d}_{21}, \mathbf{d}_{22}\} \\
\mathcal{C}_3 &= \{\mathbf{d}_7, \mathbf{d}_{12}, \mathbf{d}_{13}, \mathbf{d}_{14}, \mathbf{d}_{16}, \mathbf{d}_{23}, \mathbf{d}_{24}\}
\end{aligned}$$

Extensions and Variations

Choosing Initial Cluster Centroids

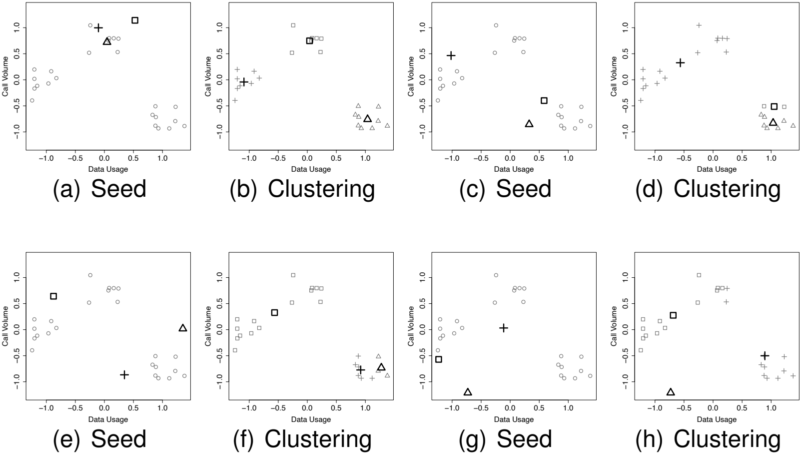


Figure 4: (a)–(h) Different clusterings (all with $k = 3$) that can be found for the mobile phone customer dataset given in Table 1[10] when different initial cluster centroids are used.

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

Require: a dataset $\mathcal{D}$ containing n training instances, $\mathbf{d}_1, \ldots, \mathbf{d}_n$
Require: k, the number of cluster centroids to find
Require: a distance measure $Dist$ to compare instances to cluster centroids

```
choose d_i randomly (following a uniform distribution) from D to be the
position of the initial centroid, c_1, of the first cluster, C_1
for cluster C_j in C_2 to C_k do
    for each instance, d_i, in D let Dist(d_i) be the distance between d_i
and its nearest cluster centroid
    calculate a selection weight for each instance, d_i, in D as
Dist(d_i)^2 / sum_{p=1}^{n} Dist(d_p)^2
    choose d_i as the position of cluster centroid, c_j, for cluster C_j
randomly following a distribution based on the selection weights
end for
proceed with k-means as normal using {c_1, ..., c_k} as the initial
centroids.
```

Step 4 selection weight:

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

Choosing Initial Cluster Centroids

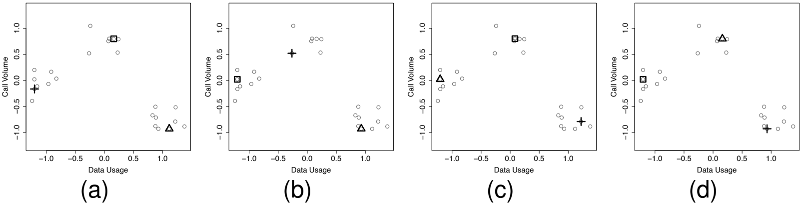


Figure 5: (a)–(d) Initial centroids chosen using the k-means++ approach (all with $k = 3$) for the mobile phone customer dataset given in Table 1[10].

Evaluating Clustering

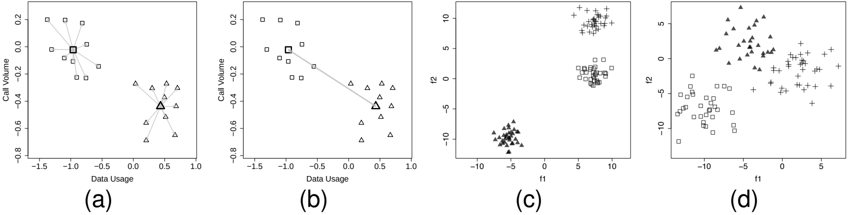


Figure 6: (a) Intra-cluster distance; (b) inter-cluster distance; (c) a *good* clustering; and (d) a *bad* clustering.

Evaluating Clustering

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

Evaluating Clustering

$$\begin{array}{cc} & \begin{array}{ccccccc} \mathbf{d}_2 & \mathbf{d}_3 & \mathbf{d}_5 & \mathbf{d}_6 & \mathbf{d}_{11} & \mathbf{d}_{19} & \mathbf{d}_{20} \end{array} \\ \mathbf{d}_1 & \left[\begin{array}{ccccccc} 0.45 & 0.30 & 0.45 & 0.50 & 0.28 & 0.44 & 0.39 \end{array}\right] \end{array}$$

$$\begin{array}{cc} & \begin{array}{ccccccccc} \mathbf{d}_4 & \mathbf{d}_8 & \mathbf{d}_9 & \mathbf{d}_{10} & \mathbf{d}_{15} & \mathbf{d}_{17} & \mathbf{d}_{18} & \mathbf{d}_{21} & \mathbf{d}_{22} \end{array} \\ \mathbf{d}_1 & \left[\begin{array}{ccccccccc} 2.03 & 1.88 & 2.09 & 1.94 & 1.83 & 2.11 & 1.95 & 1.80 & 1.74 \end{array}\right] \end{array}$$

$$\begin{array}{cc} & \begin{array}{ccccccc} \mathbf{d}_7 & \mathbf{d}_{12} & \mathbf{d}_{13} & \mathbf{d}_{14} & \mathbf{d}_{16} & \mathbf{d}_{23} & \mathbf{d}_{24} \end{array} \\ \mathbf{d}_1 & \left[\begin{array}{ccccccc} 1.16 & 1.59 & 1.38 & 1.64 & 1.64 & 1.26 & 0.95 \end{array}\right] \end{array}$$

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, $\mathbf{d}_1, \ldots, \mathbf{d}_n$
Require: a clustering $\mathcal{C}$ of dataset $\mathcal{D}$ into k clusters, $\mathcal{C}_1, \ldots, \mathcal{C}_k$
Require: a distance measure, $Dist$, to compare distances between instances

for each instance $\mathbf{d}_i$ in $\mathcal{D}$ **do**
    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*)
    calculate the average distance between instance $\mathbf{d}_i$ and the members of each of the other clusters $\mathcal{C} \setminus \mathcal{C}_j$
    let $b(i)$ be the lowest average distance between instance $\mathbf{d}_i$ and any other cluster (*average inter-cluster distance*)
    calculate the silhouette index for $\mathbf{d}_i$ as

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3}$$

end for
calculate final silhouette for the clustering as $s = \frac{1}{n} \sum_{i=1}^{n} s(i)$

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	$a(i)$	$b(i)$	$s(i)$
1	$\mathcal{C}_1$	$\mathcal{C}_3$	0.401	1.374	0.708
2	$\mathcal{C}_1$	$\mathcal{C}_3$	0.695	1.811	0.616
3	$\mathcal{C}_1$	$\mathcal{C}_3$	0.503	1.644	0.694
4	$\mathcal{C}_2$	$\mathcal{C}_3$	0.484	1.628	0.703
5	$\mathcal{C}_1$	$\mathcal{C}_3$	0.387	1.232	0.686
6	$\mathcal{C}_1$	$\mathcal{C}_3$	0.445	0.970	0.541
7	$\mathcal{C}_3$	$\mathcal{C}_1$	0.452	1.056	0.572
8	$\mathcal{C}_2$	$\mathcal{C}_3$	0.599	1.364	0.561
9	$\mathcal{C}_2$	$\mathcal{C}_3$	0.470	1.768	0.734
10	$\mathcal{C}_2$	$\mathcal{C}_3$	0.504	1.978	0.745
11	$\mathcal{C}_1$	$\mathcal{C}_3$	0.327	1.223	0.732
12	$\mathcal{C}_3$	$\mathcal{C}_1$	0.433	1.537	0.719

ID	Cluster	Nearest Cluster	$a(i)$	$b(i)$	$s(i)$
13	$\mathcal{C}_3$	$\mathcal{C}_1$	0.5136	1.3592	0.6221
14	$\mathcal{C}_3$	$\mathcal{C}_1$	0.4349	1.5738	0.7236
15	$\mathcal{C}_2$	$\mathcal{C}_3$	0.5776	1.3480	0.5715
16	$\mathcal{C}_3$	$\mathcal{C}_1$	0.4955	1.5409	0.6784
17	$\mathcal{C}_2$	$\mathcal{C}_1$	0.7369	2.2757	0.6762
18	$\mathcal{C}_2$	$\mathcal{C}_3$	0.4312	1.8473	0.7666
19	$\mathcal{C}_1$	$\mathcal{C}_3$	0.3711	1.1682	0.6823
20	$\mathcal{C}_1$	$\mathcal{C}_3$	0.4334	1.0006	0.5669
21	$\mathcal{C}_2$	$\mathcal{C}_1$	0.6520	1.9710	0.6692
22	$\mathcal{C}_2$	$\mathcal{C}_3$	0.4504	1.5457	0.7086
23	$\mathcal{C}_3$	$\mathcal{C}_1$	0.3954	1.1654	0.6607
24	$\mathcal{C}_3$	$\mathcal{C}_1$	0.5339	0.8880	0.3988

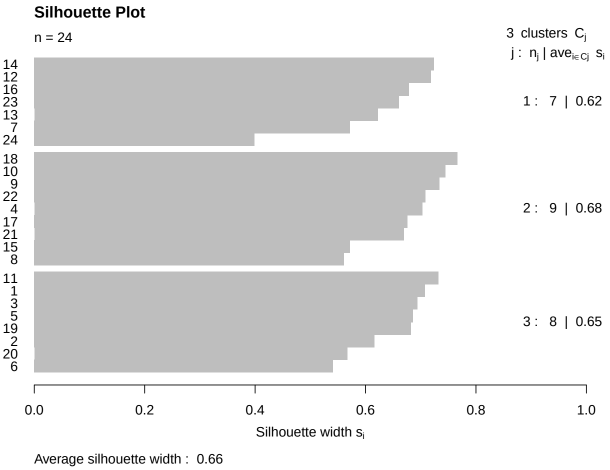
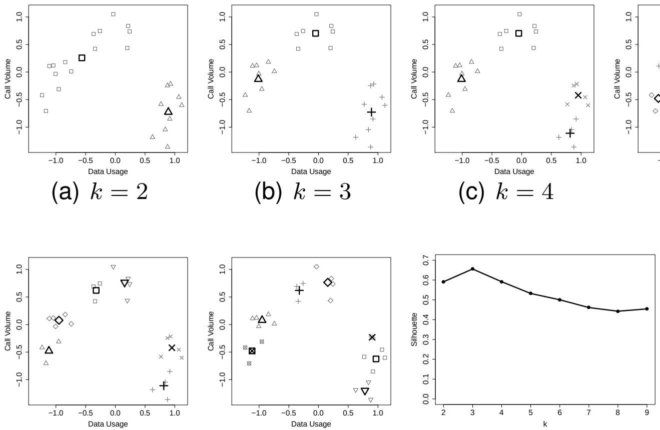


Figure 7: The silhouette plot for the final clustering of the mobile phone customer dataset (Table 1[10]) found using the *k*-means algorithm (with $k = 3$).

Choosing the Number of Clusters



(a) $k = 2$ (b) $k = 3$ (c) $k = 4$ (d) $k = 5$

(e) $k = 6$ (f) $k = 7$ (g) Silhouette summary

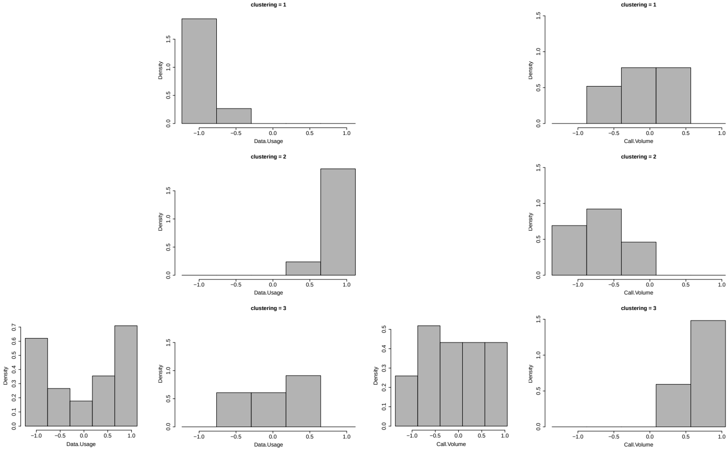
Figure 8: (a)–(f) Different clusterings found for the mobile phone customer dataset in Table 1[10] for values of k in $(2, 9)$. (g) shows the silhouette for each clustering.

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	Min.	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max	Std. Dev.
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

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 ($k = 3$).

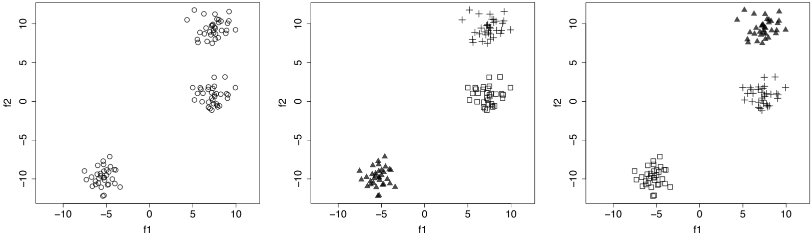
$\mathcal{C}_1$		$\mathcal{C}_2$		$\mathcal{C}_3$	
Feature	Info. Gain	Feature	Info. Gain	Feature	Info. Gain
DATA.USAGE	0.9183	DATA.USAGE	0.9544	CALL.VOLUME	0.8709
CALL.VOLUME	0.2117	CALL.VOLUME	0.5488	DATA.USAGE	0.2479



(a) DATA USAGE

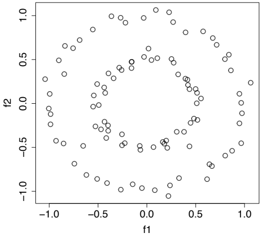
(b) CALL VOLUME

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 ($k = 3$).

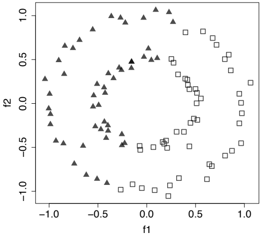


(a) Blobs Dataset (b) k-means (c) AHC

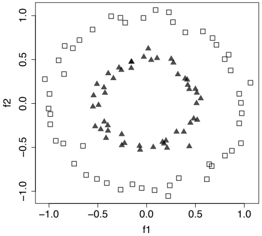
Figure 10: (a)–(i) A plot of the *blobs*, *circles*, and *half-moons* datasets and the clusterings achieved by the k-means clustering and agglomerative hierarchical clustering algorithms (where k is set to 3, 2, and 2, respectively).



(d) Circles Dataset

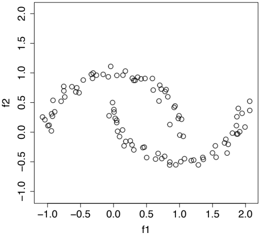


(e) k-means

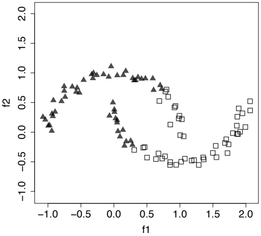


(f) AHC

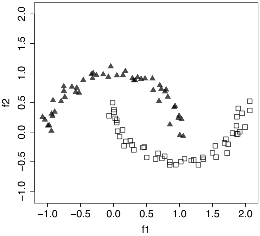
Figure 11: (a)–(i) A plot of the *blobs*, *circles*, and *half-moons* datasets and the clusterings achieved by the *k*-means clustering and agglomerative hierarchical clustering algorithms (where k is set to 3, 2, and 2, respectively).



(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, $\mathbf{d}_1, \ldots, \mathbf{d}_n$
Require: a distance measure, $Dist$, to compare distances between instances
Require: a linkage method, $\mathcal{L}$, to compare distances between clusters

```
initialize the hierarchy level, h = 1
divide D into a set of n disjoint clusters, C = {C_1, ..., C_n}, with one instance in each cluster
repeat
    using distance measure Dist and linkage method L, find the nearest pair of clusters, C_i and C_j, in the current clustering
    merge C_i and C_j to form a new cluster C_{n+h}
    remove the old clusters from the clustering: C ← C \ {C_i, C_j}
    add the new cluster to the clustering: C ← C ∪ C_{n+h}
    h ← h + 1
until all the instances join into a single cluster
```

Agglomerative Hierarchical Clustering

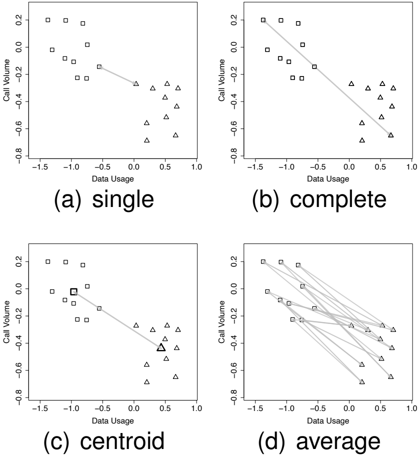


Figure 13: (a)–(d) Different linkage methods that can be used to compare the distances between clusters in agglomerative hierarchical clustering. (Arrows for only some indicative distances are shown in the average linkage diagram (d).)

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	$\mathcal{C}_{10}$	d_{11}	d_5	d_{19}	d_{24}	d_7	d_{23}
d_4	0.00							
$\mathcal{C}_{10}$	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	0.00

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

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

	d_4	$\mathcal{C}_{13}$	$\mathcal{C}_{11}$	d_{24}	$\mathcal{C}_{12}$
d_4	0.00				
$\mathcal{C}_{13}$	0.28	0.00			
$\mathcal{C}_{11}$	2.19	0.16	0.00		
d_{24}	1.66	0.81	0.76	0.00	
$\mathcal{C}_{12}$	1.79	0.97	0.89	0.27	0.00

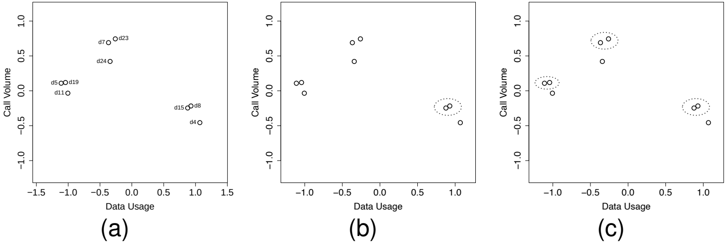
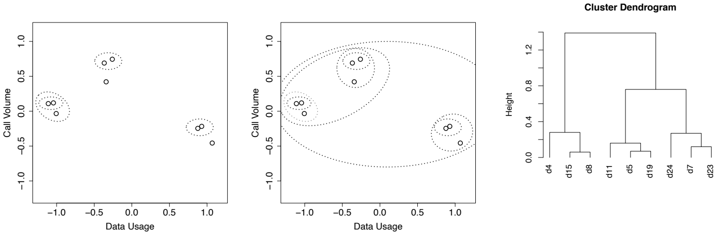
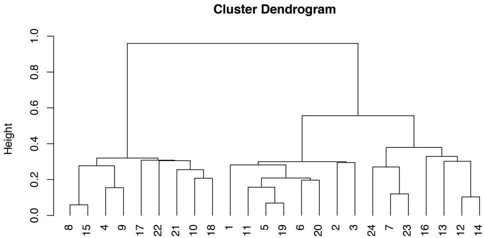


Figure 14: (a) A plot of a reduced version of the mobile phone customer dataset given in Table 1[10]. (b) At the first iteration of the AHC algorithm the first pair of instances is combined into a cluster, $\mathcal{C}_{10}$. (c) After three iterations of the AHC algorithm, three pairs of instances have 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, $\mathcal{C}_{13}$.



(d) (e) (f)

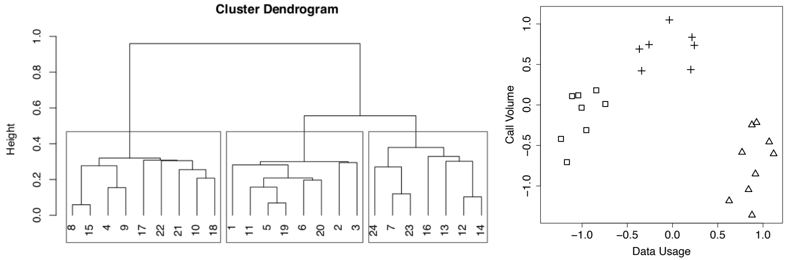
Figure 15: (a) A plot of a reduced version of the mobile phone customer dataset given in Table 1[10]. (b) At the first iteration of the AHC algorithm the first pair of instances is combined into a cluster, $\mathcal{C}_{10}$. (c) After three iterations of the AHC algorithm, three pairs of instances have 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, $\mathcal{C}_{13}$.



(a) AHC result

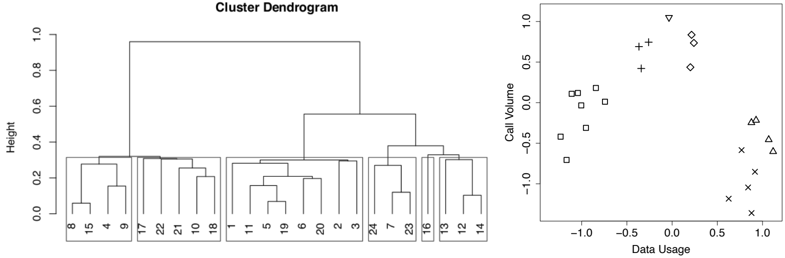
Figure 16: (a) A plot of the hierarchical grouping of the instances in the mobile phone customer dataset from Table 1[10] found by the AHC algorithm (using Euclidean distance and single linkage). (b) The clustering returned when the tree is cut at $k = 3$. (c) The clustering returned when the tree is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

Agglomerative Hierarchical Clustering



(b) Clustering cut at $k = 3$

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 $k = 6$.



(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

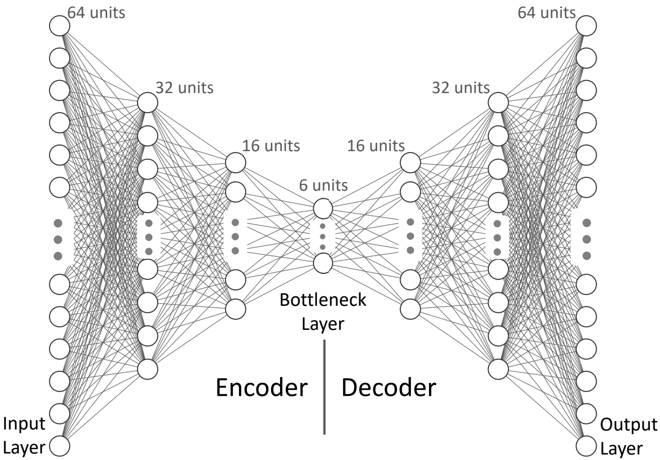


Figure 19: The architecture of an auto-encoder network made up of an encoder and a decoder connected by a bottleneck layer.

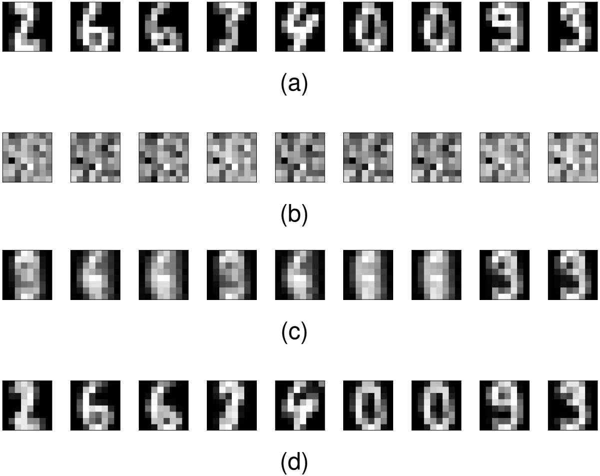


Figure 20: (a) A selection of images from the handwritten digits dataset; (b) image reconstructions generated by the auto-encoder network before training; (c) image reconstructions generated by the auto-encoder network after minimal training (10 epochs); and (d) image reconstructions generated by the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

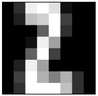
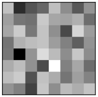
Training	Image	Pixel Values	Error
Original		$\begin{bmatrix} 0.00 & 0.19 & 0.94 & 1.00 & 0.88 & 0.06 & 0.00 & 0.00 \\ 0.00 & 0.12 & 0.75 & 0.81 & 1.00 & 0.25 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.38 & 1.00 & 0.19 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.06 & 0.94 & 0.62 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.38 & 1.00 & 0.25 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.12 & 0.94 & 0.62 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.25 & 1.00 & 0.69 & 0.50 & 0.69 & 0.19 & 0.00 \\ 0.00 & 0.19 & 1.00 & 1.00 & 1.00 & 0.75 & 0.19 & 0.00 \end{bmatrix}$	
0 Epochs		$\begin{bmatrix} 0.51 & 0.48 & 0.49 & 0.49 & 0.51 & 0.50 & 0.49 & 0.50 \\ 0.51 & 0.50 & 0.49 & 0.52 & 0.51 & 0.51 & 0.51 & 0.49 \\ 0.51 & 0.52 & 0.50 & 0.51 & 0.50 & 0.49 & 0.52 & 0.50 \\ 0.50 & 0.51 & 0.51 & 0.51 & 0.50 & 0.50 & 0.49 & 0.50 \\ 0.50 & 0.47 & 0.50 & 0.52 & 0.51 & 0.50 & 0.52 & 0.50 \\ 0.51 & 0.51 & 0.50 & 0.49 & 0.53 & 0.50 & 0.51 & 0.49 \\ 0.51 & 0.52 & 0.49 & 0.51 & 0.51 & 0.50 & 0.51 & 0.50 \\ 0.50 & 0.50 & 0.48 & 0.51 & 0.50 & 0.51 & 0.51 & 0.51 \end{bmatrix}$	0.1876

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.

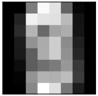
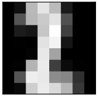
Training	Image	Pixel Values	Error
10 Epochs		$\begin{bmatrix} 0.00 & 0.00 & 0.42 & 0.83 & 0.76 & 0.33 & 0.03 & 0.00 \\ 0.01 & 0.12 & 0.80 & 0.76 & 0.74 & 0.62 & 0.06 & 0.00 \\ 0.00 & 0.15 & 0.60 & 0.35 & 0.54 & 0.58 & 0.07 & 0.00 \\ 0.00 & 0.09 & 0.49 & 0.57 & 0.68 & 0.46 & 0.11 & 0.00 \\ 0.00 & 0.08 & 0.31 & 0.57 & 0.68 & 0.48 & 0.13 & 0.01 \\ 0.00 & 0.05 & 0.31 & 0.37 & 0.41 & 0.51 & 0.19 & 0.00 \\ 0.00 & 0.02 & 0.49 & 0.59 & 0.59 & 0.63 & 0.21 & 0.01 \\ 0.00 & 0.01 & 0.45 & 0.85 & 0.74 & 0.39 & 0.09 & 0.01 \end{bmatrix}$	0.0685
$1{,}000$ Epochs		$\begin{bmatrix} 0.00 & 0.06 & 0.71 & 0.87 & 0.71 & 0.10 & 0.00 & 0.00 \\ 0.00 & 0.32 & 0.70 & 0.77 & 0.86 & 0.30 & 0.00 & 0.00 \\ 0.00 & 0.11 & 0.09 & 0.75 & 0.97 & 0.24 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.86 & 0.93 & 0.08 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.02 & 0.88 & 0.62 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.01 & 0.68 & 0.89 & 0.24 & 0.04 & 0.01 & 0.00 \\ 0.00 & 0.32 & 0.91 & 0.89 & 0.53 & 0.51 & 0.19 & 0.00 \\ 0.00 & 0.03 & 0.78 & 0.89 & 0.83 & 0.69 & 0.21 & 0.00 \end{bmatrix}$	0.0179

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.

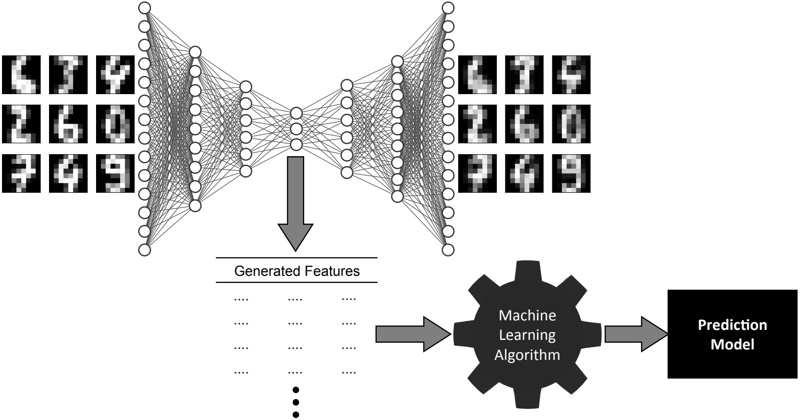


Figure 23: The process of using an unsupervised auto-encoder network to generate a feature representation used to train a supervised model.

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.

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.