Advanced Data Analysis

DATA 71200

Class 9: Unsupervised Learning (Clustering)

Schedule

22-Jun	**Unsupervised Learning (Clustering)** **Ethics (Part Two)**
1-Jul	*Project 3 Due*
8-Jul	*Final Paper Due* *Data Camp Assignments Due*

Project 3 (Due July 1)

Project 3

To **submit** the assignment, submit a link to your project Jupyter notebook on Blackboard.

The **goal** for this assignment is two apply different types of unsupervised learning techniques on the dataset you created in Project 1.

Step 1: Load your data, including testing/training split from Project 1.

- Your testing and training split should be balanced
- Your data should be clean and missing data should be addressed
- All appropriate variables are converted to categorical variables (as ordinal or one hot)
- Any necessary feature scaling should be performed
- YOU SHOULD ONLY WORK ON YOUR TRAINING SET

Project 3 (Due July 1)

Step 2: PCA for feature selection

- Show how many features do you need to retain to capture 95% of the variance
- Evaluate whether this improves your best-performing model from Project 2

Step 3: Apply 3-types of clustering on your data and visualize the output of each *both with and without PCA run on it first*. Calculate both ARI and Silhouette Coefficient for all six of the combinations.

- k-Means (use an elbow visualization to determine the optimal numbers of clusters)
- Aggolmerate/Hierarchical
- DBSCAN

If your data from projects 1 and 2 really doesn't lend itself to clustering,
you can use the breast_cancer dataset from scikit-learn.

Still submit your attempts on your own data in the notebook.

Tip: You should make notes on what worked well and what didn't. Such notes will be useful when you write up the paper for your final project.

Project 3 Rubric

	Missing	Fair	Good	Excellent
Data	**0** (0.00%)	**2** (13.33333%) One of: Balanced testing/training split. Only using training set. Scaling and encoding done, where appropriate.	**2.5** (16.66666%) Two of: Balanced testing/training split. Only using training set. Scaling and encoding done, where appropriate.	**3** (20.00%) Balanced testing/training split. Only using training set. Scaling and encoding done, where appropriate.
PCA (95% variance)	**0** (0.00%)	**0.5** (3.33333%) PCA performed but captures a different amount of variance captured.	**0** (0.00%)	**1** (6.66666%) PCA captures 95% of variance
PCA on supervised learning	**0** (0.00%)	**0** (0.00%)	**0** (0.00%) PCA feature selection performed with some issue as a pre-processing step for supervised learning.	**1** (6.66666%) PCA feature selection accurately performed as a pre-processing step for supervised learning.
k-Means	**0** (0.00%)	**2** (13.33333%) Run and neither visualized or with and without PCA	**2.5** (16.66666%) Run and either visualized or with and without PCA	**3** (20.00%) Run and visualized with and without PCA
k-Means elbow visualization	**0** (0.00%)	**0.5** (3.33333%) Run and results not used to select number of clusters	**0** (0.00%)	**1** (6.66666%) Run and results used to select number of clusters
Agglomerative/Hierarchical	**0** (0.00%)	**2** (13.33333%) Run and neither visualized or with and without PCA	**2.5** (16.66666%) Run and either visualized or with and without PCA	**3** (20.00%) Run and visualized with and without PCA
DBSCAN	**0** (0.00%)	**2** (13.33333%) Run and neither visualized or with and without PCA	**2.5** (16.66666%) Run and either visualized or with and without PCA	**3** (20.00%) Run and visualized with and without PCA

Clustering

- **Algorithms that assign data points to groups (especially for unlabeled data)**
  - In the absence of labels, evaluation is challenging
    - Often performed through visualization
- **Useful for**
  - Exploratory data analysis
  - Pre-processing data

k-Means

- ***k* - number of clusters specified**
- **Finds cluster centers through an iterative process**
  - Assign data points to cluster with nearest cluster center
    - Initialized randomly for the first iteration
  - Update the cluster center with the assigned data points
  - Repeat until no updates are needed
- **Boundaries are determined by placement of cluster centers**

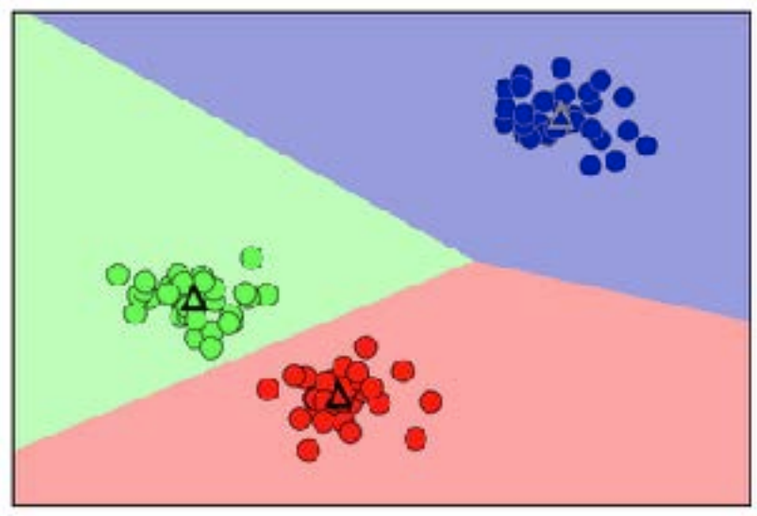


Figure 3-24. Cluster centers and cluster boundaries found by the k-means algorithm

Muller, Andreas C. and Sarah Guido. (2016). Introduction to Machine Learning with Python, O'Reilly Media, Inc.

k-Means

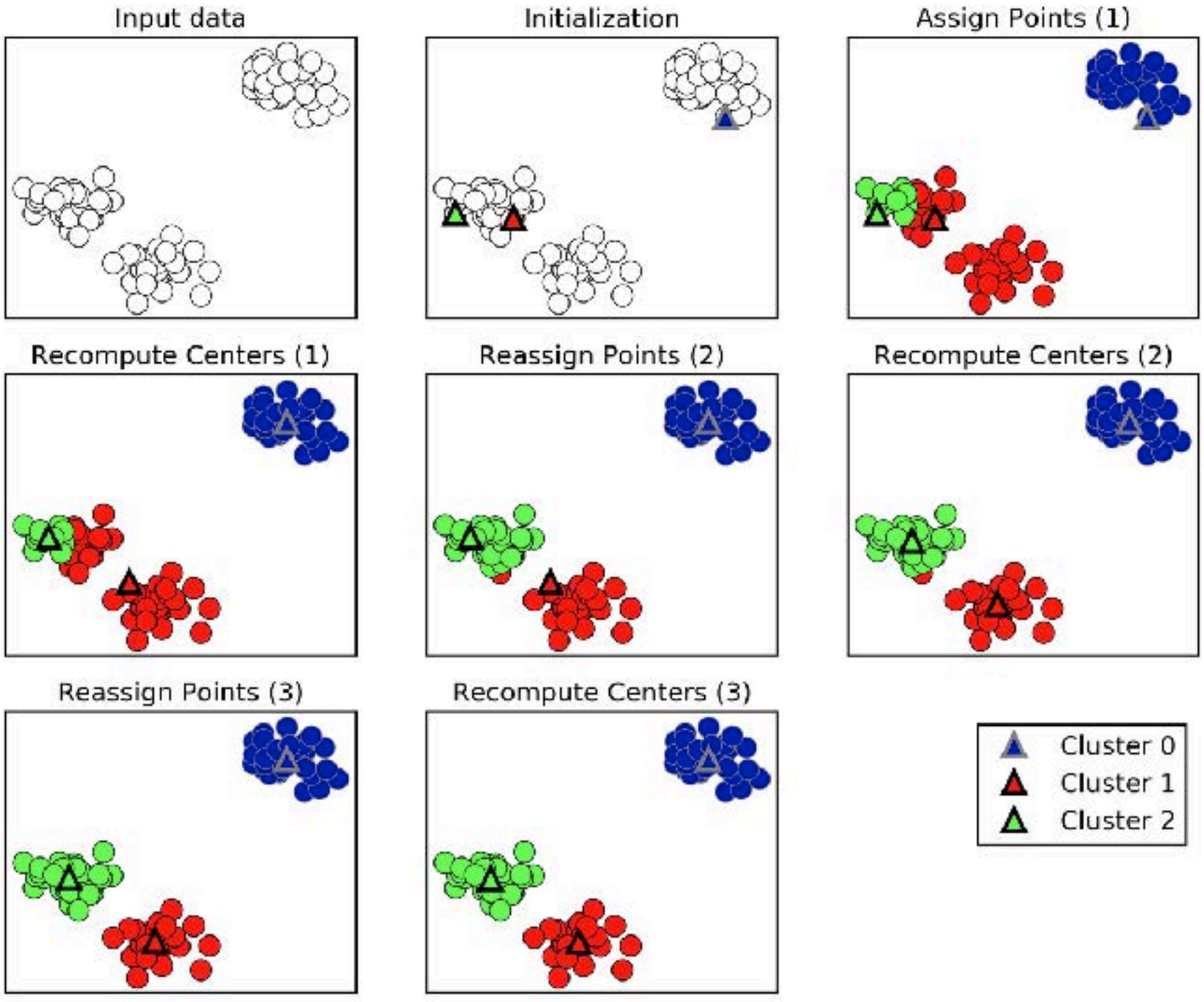


Figure 3-23. Input data and three steps of the k-means algorithm

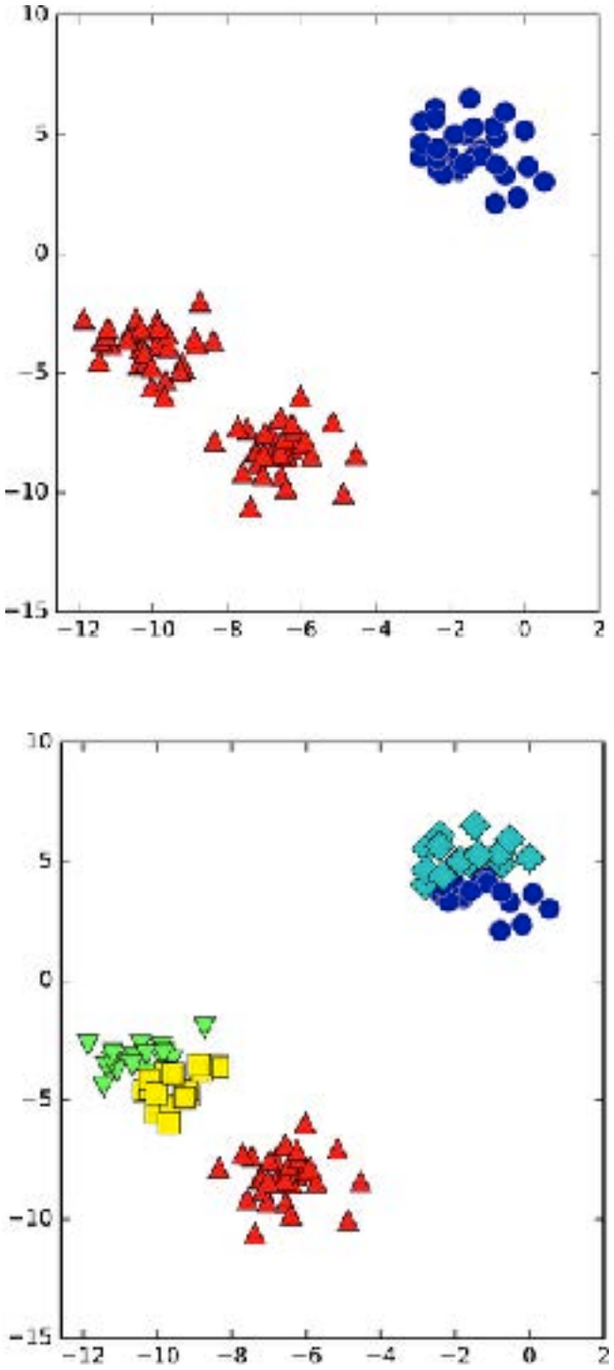


Figure 3-26. Cluster assignments found by k-means using two clusters (top) and five clusters (bottom)

Muller, Andreas C. and Sarah Guido. (2016). Introduction to Machine Learning with Python, O'Reilly Media, Inc.

k-Means

- **Assumes the classes have the same width/diameter**
- **This causes issues with non-spherical clusters or clusters with complex shapes**

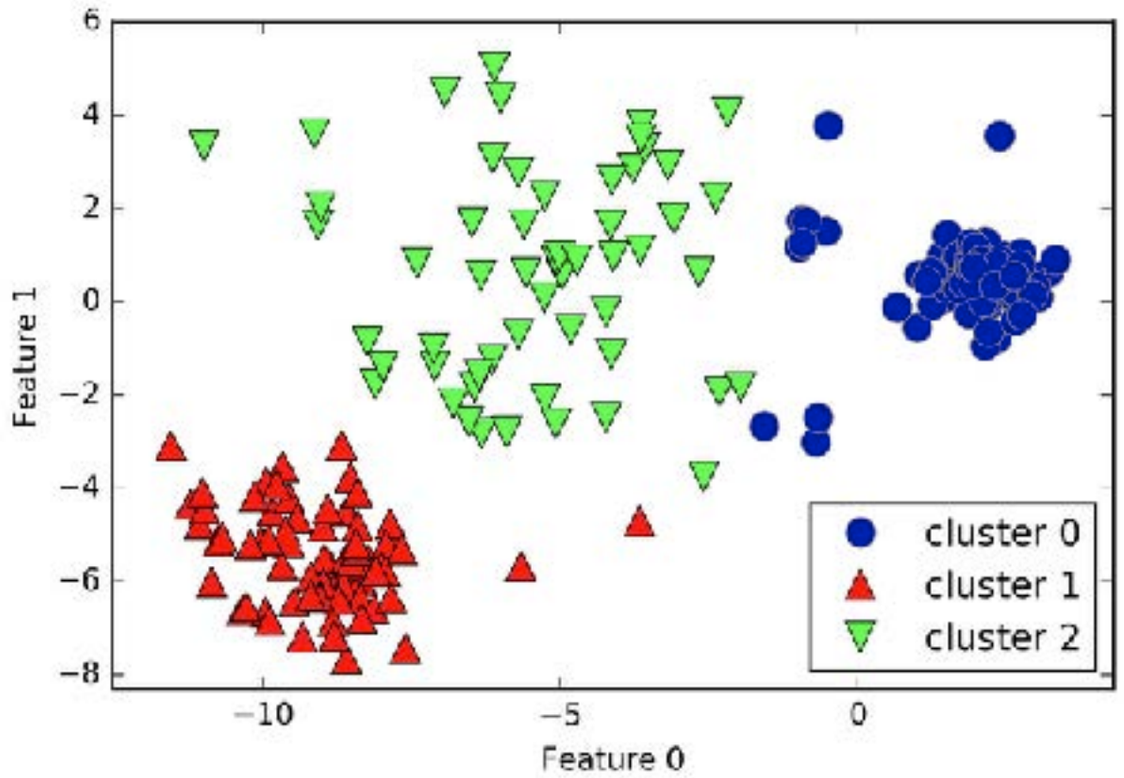


Figure 3-27. Cluster assignments found by k-means when clusters have different densities

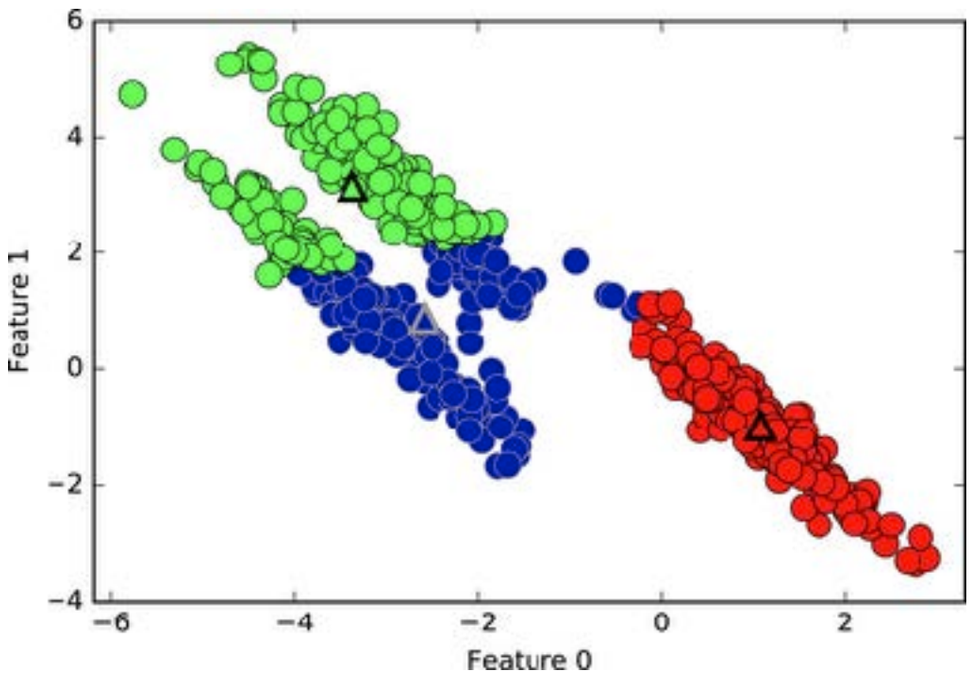


Figure 3-28. k-means fails to identify nonspherical clusters

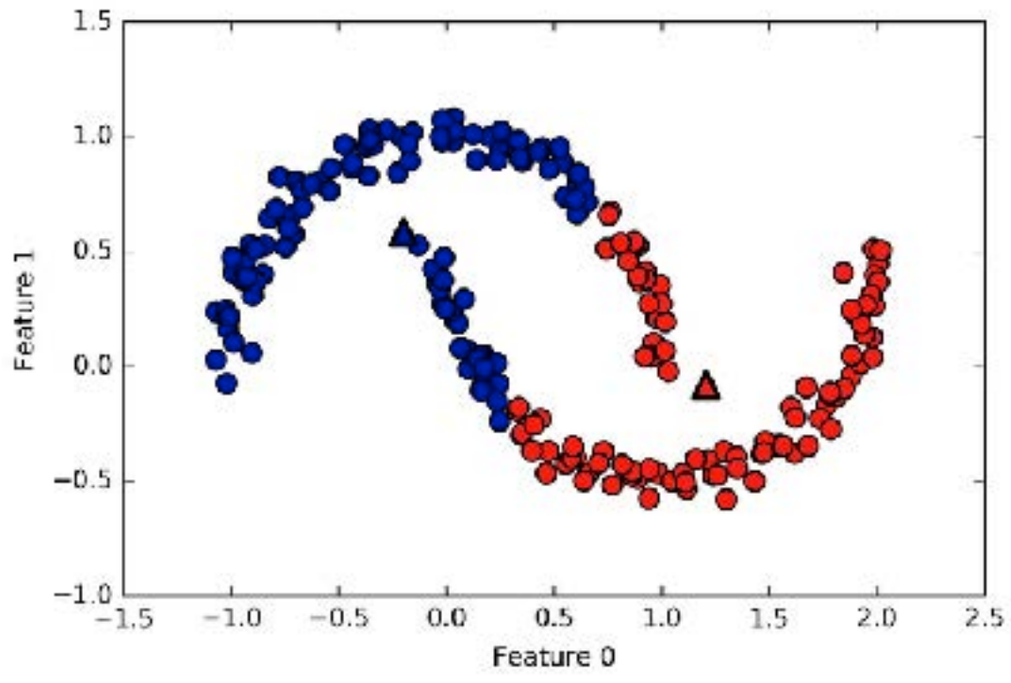


Figure 3-29. k-means fails to identify clusters with complex shapes

Muller, Andreas C. and Sarah Guido. (2016). Introduction to Machine Learning with Python, O'Reilly Media, Inc.

k-Means

- **The constant-width limitation can be partially overcome with a larger number of clusters**

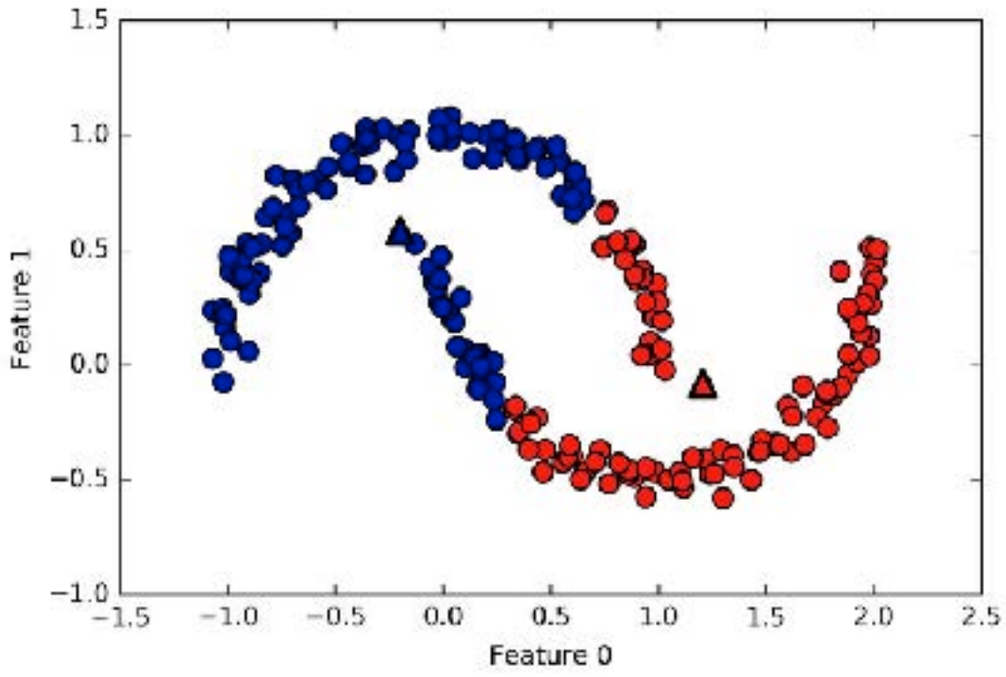


Figure 3-29. k-means fails to identify clusters with complex shapes

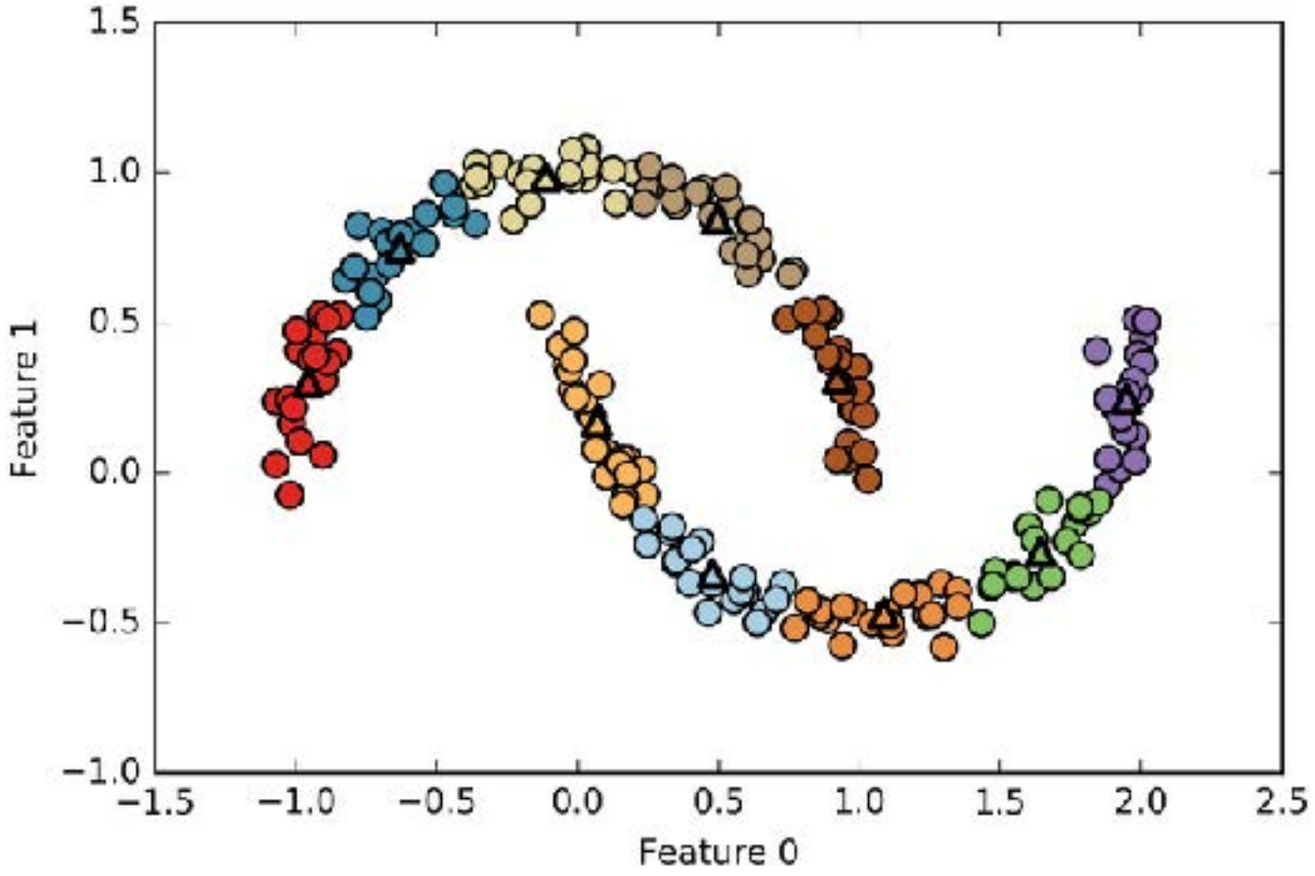


Figure 3-32. Using many k-means clusters to cover the variation in a complex dataset

Muller, Andreas C. and Sarah Guido. (2016). Introduction to Machine Learning with Python, O'Reilly Media, Inc.

k-Means

- **Best Practices**
  - Can run in batches on very large datasets
- **Strengths**
  - Easy to understand
  - Runs relatively quickly
- **Weaknesses**
  - Based on random initialization
  - Need to specify the number of clusters
  - Clusters have consistent widths and shapes

Agglomerative Clustering

- Starts by creating a cluster for each point
- Then amalgamates nearest clusters based on linkage criteria until the stopping criteria is reached

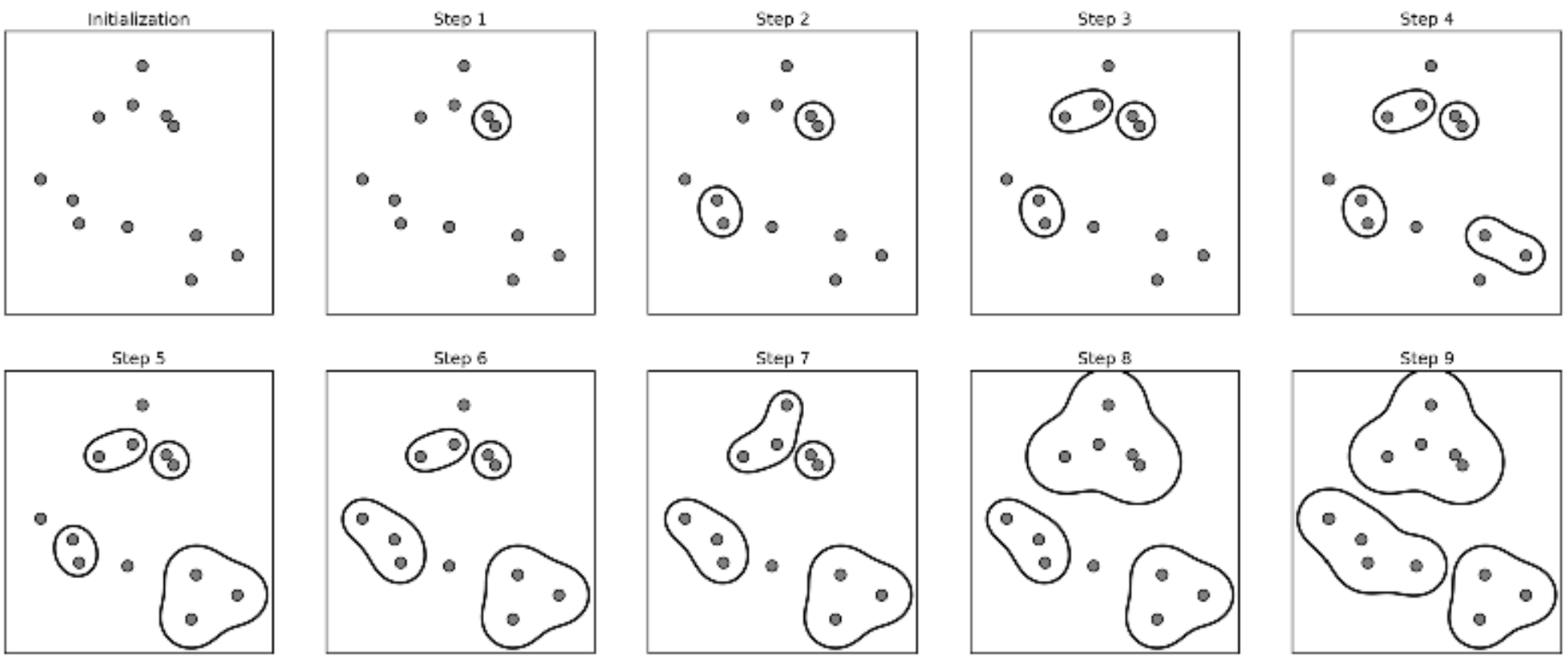
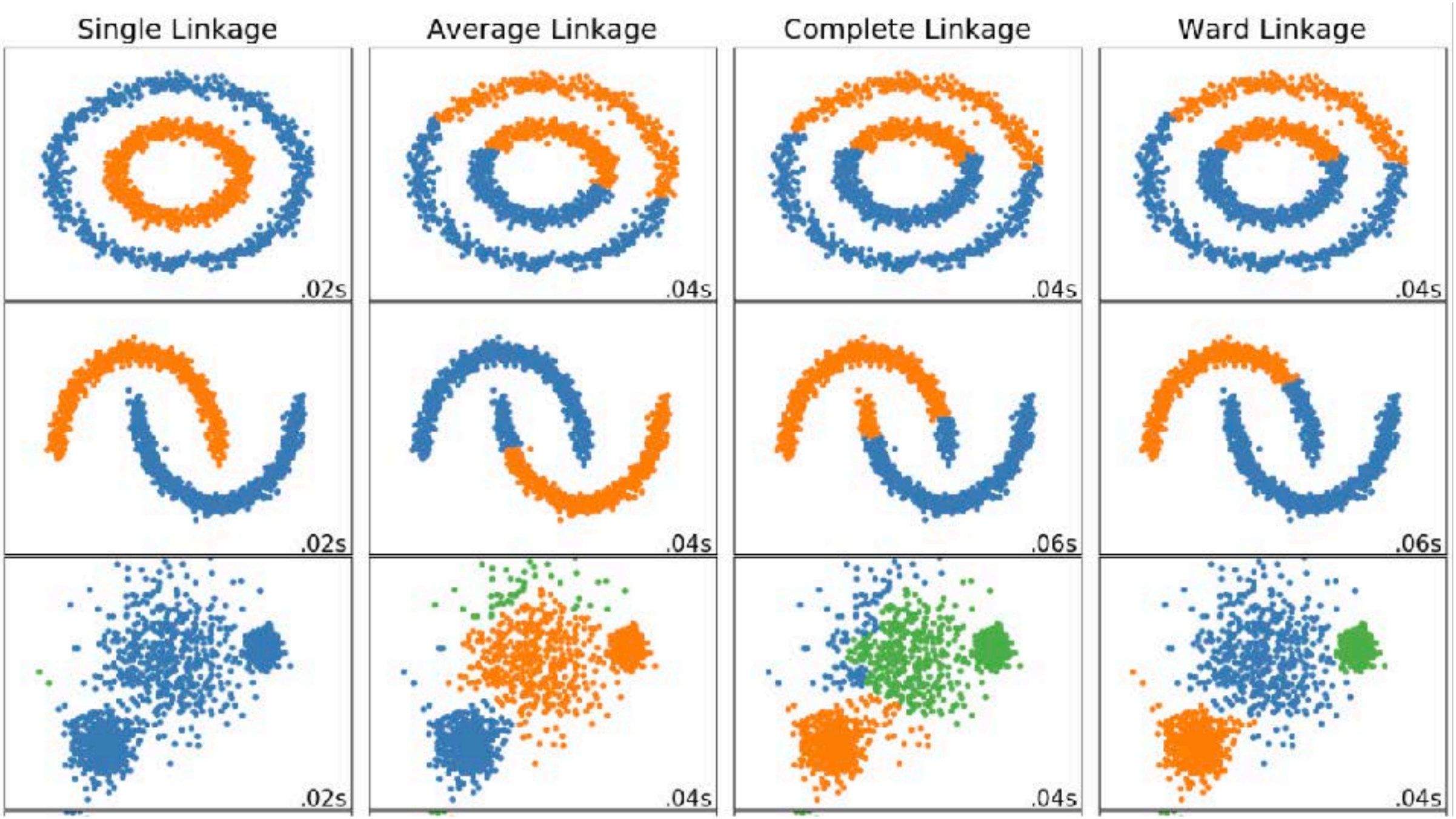


Figure 3-33. Agglomerative clustering iteratively joins the two closest clusters

Muller, Andreas C. and Sarah Guido. (2016). Introduction to Machine Learning with Python, O'Reilly Media, Inc.

Agglomerative Clustering



uses the **minimum** of the distances between all observations of the two sets

uses the **average** of the distances of each observation of the two sets

uses the **maximum** distances between all observations of the two sets

minimizes the **variance** of the clusters being merged

https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

Agglomerative Clustering

- **Looking at all possible clusters simultaneously provides information about the hierarchical relationship of the clusters**
- **Dendrograms allow for visualization of multidimensional datasets, also providing information about cluster distance**

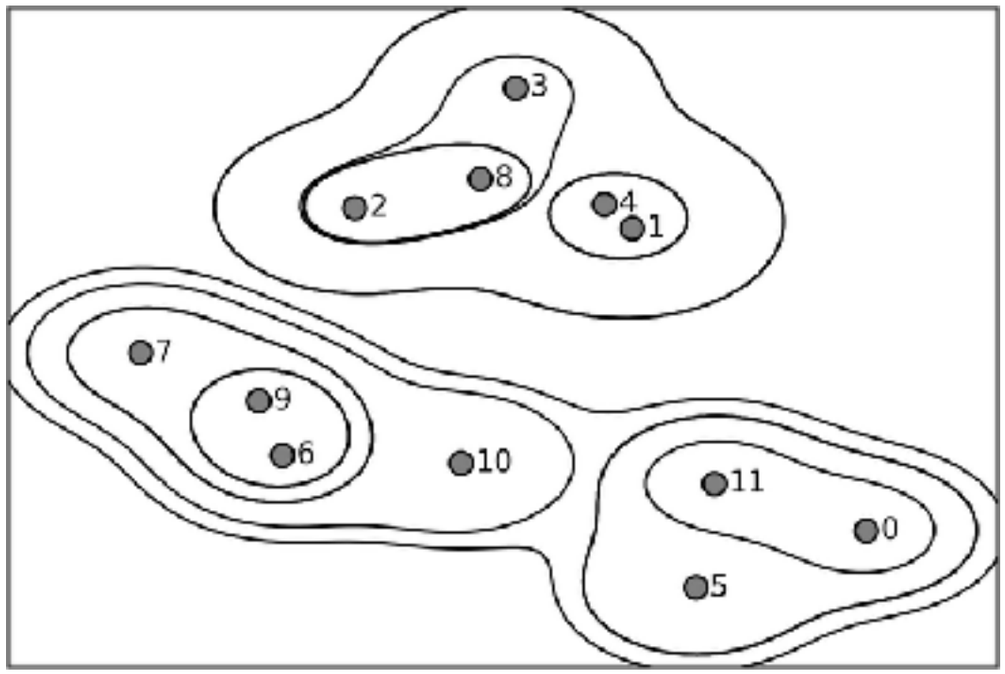


Figure 3-35. Hierarchical cluster assignment (shown as lines) generated with agglomerative clustering, with numbered data points (cf. Figure 3-36)

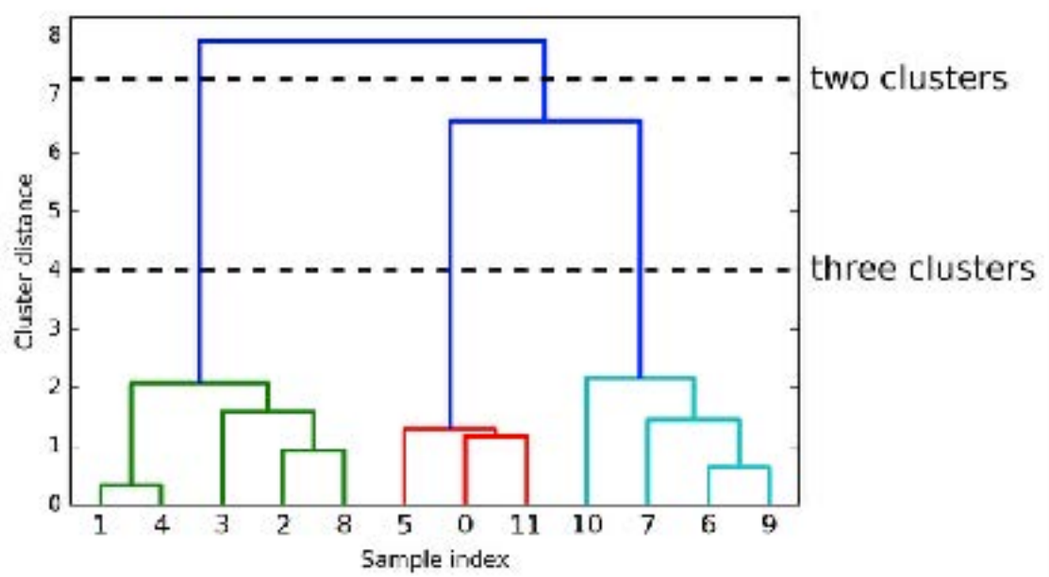


Figure 3-36. Dendrogram of the clustering shown in Figure 3-35 with lines indicating splits into two and three clusters

Muller, Andreas C. and Sarah Guido. (2016). Introduction to Machine Learning with Python, O'Reilly Media, Inc.

Agglomerative Clustering

- **Parameters**
  - Linkage criteria: ward, single, average, complete
  - Stopping criteria: number of clusters
- **Strengths**
  - Easy to understand/visualize
- **Weaknesses**
  - Not able to make prediction on new data
  - In scikit-learn you need to specify the number of clusters

DBSCAN

- **Density-based spatial clustering of applications with noise**
- **Do not need to specify the number of clusters**
- **Attempts to distinguish between densely and sparsely populated areas of the data space**
  - Core points - cluster centers
  - Boundary points - within a cluster
  - Noise

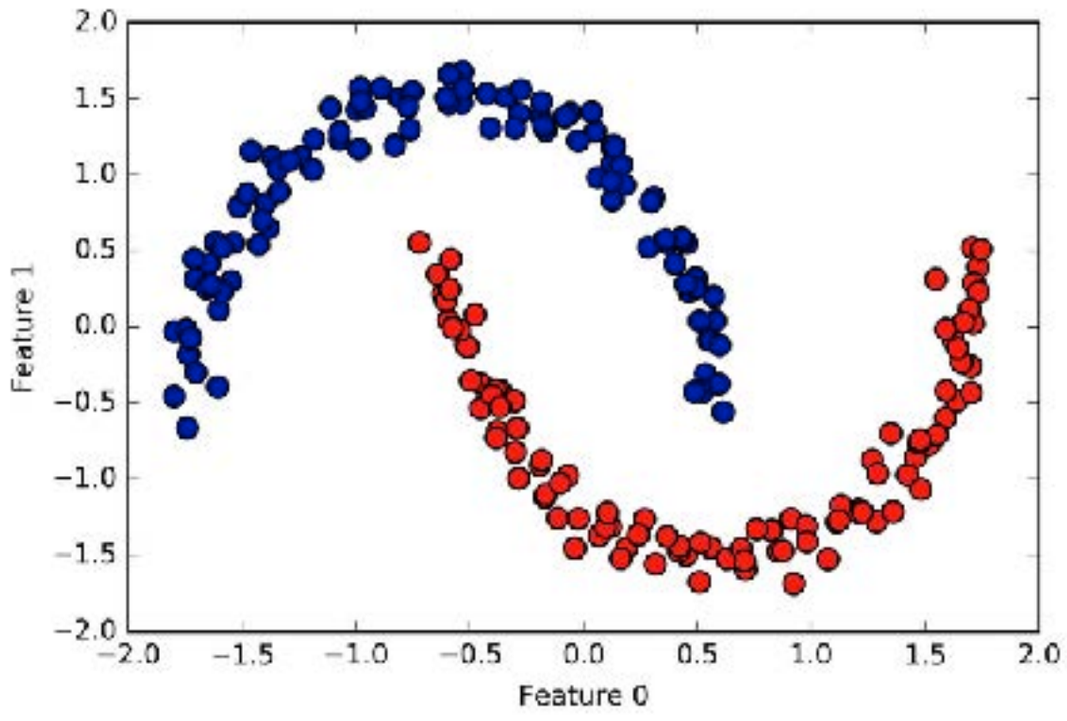


Figure 3-38. Cluster assignment found by DBSCAN using the default value of eps=0.5

Muller, Andreas C. and Sarah Guido. (2016). Introduction to Machine Learning with Python, O'Reilly Media, Inc.

DBSCAN

- Procedure (repeated until clusterable data has been addressed)
  - Select a data point and check how many other data points are within the specified distance
  - If there are as many as the specified minimum number, data point is considered a core sample
    - Data points within the minimum distance are boundary points
  - If there are multiple core samples within the specified distance, they are merged into a single cluster and their neighbors are also visited
- If points aren't clustered, they are classified as noise

DBSCAN

- Increasing eps results in more points per cluster
- Increasing min_samples results in more being classified as noise

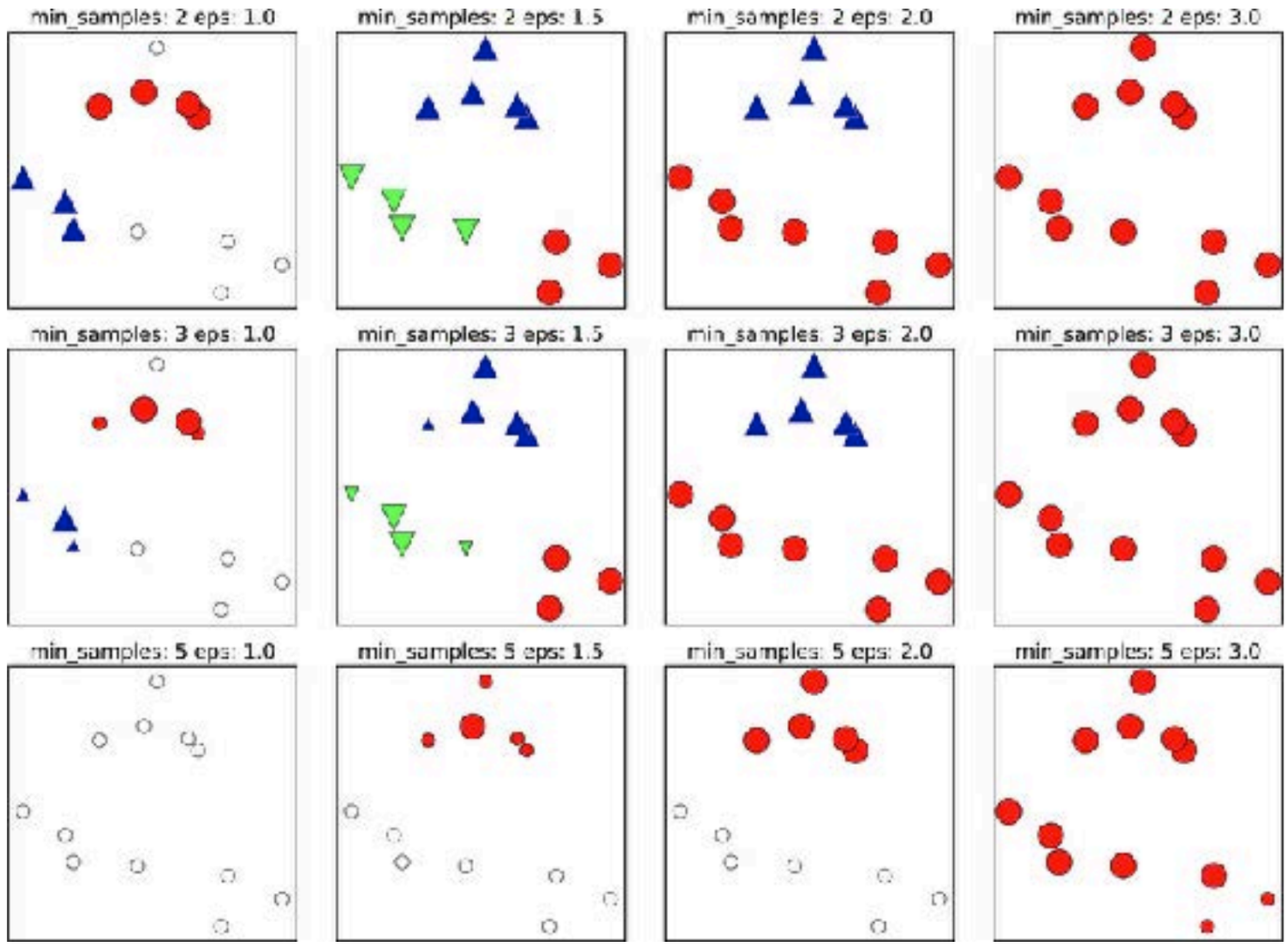


Figure 3-37. Cluster assignments found by DBSCAN with varying settings for the min_samples and eps parameters

Muller, Andreas C. and Sarah Guido. (2016). Introduction to Machine Learning with Python, O'Reilly Media, Inc.

DBSCAN

- **Best Practices**
  - Scaling data can improve clustering results with DBSCAN
- **Parameters**
  - eps - determines distance the algorithm looks for data points
  - min_samples - determines the minimum number of data points within eps distance necessary to form a cluster
- **Strengths**
  - Able to cluster complex shapes
- **Weaknesses**
  - Cluster assignment depends on order the points are visited
  - Results sensitive to the settings of min_samples and eps

Evaluating Clustering

With Ground Truth: Adjusted Rand Index (ARI)

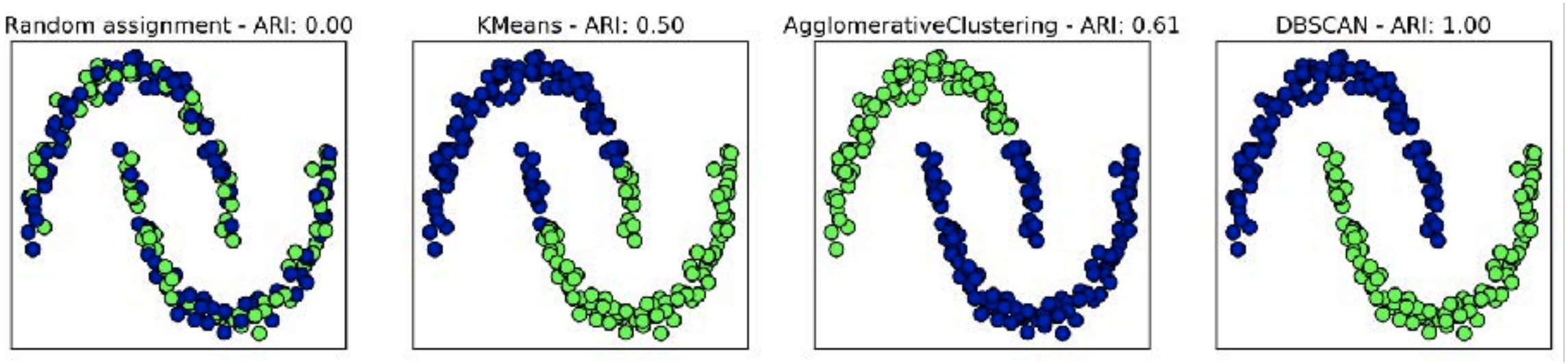


Figure 3-39. Comparing random assignment, k-means, agglomerative clustering, and DBSCAN on the two_moons dataset using the supervised ARI score

With No Ground Truth: Silhouette Coefficient

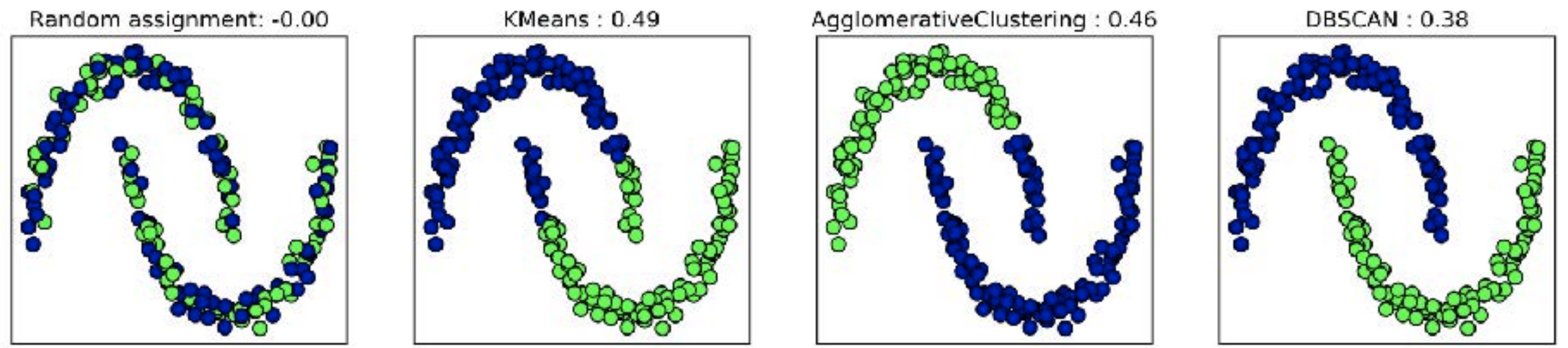


Figure 3-40. Comparing random assignment, k-means, agglomerative clustering, and DBSCAN on the two_moons dataset using the unsupervised silhouette score—the more intuitive result of DBSCAN has a lower silhouette score than the assignments found by k-means

Muller, Andreas C. and Sarah Guido. (2016). Introduction to Machine Learning with Python, O'Reilly Media, Inc.

Upcoming Work

- **Project 2 due today**
- **Videos for June 22 (password: data71200)**
  - Unsupervised Learning 2: https://vimeo.com/415374034
- **Reading for June 21**
  - Ch 3: “Unsupervised Learning” in Guido, Sarah and Andreas C. Muller. (2016). Introduction to Machine Learning with Python, O’Reilly Media, Inc. 170–211
  - West, Sarah Myers, Meredith Whittaker, and Kate Crawford. (2019). “Discriminating systems: Gender, race and power in AI.” AI Now Institute, 1–33
- **DataCamp**
  - Unsupervised Learning in Python course
- **Project 3 due July 1**

Final Paper (Due July 8)

Final Paper

A 5–8 page paper describing the work you did in projects 1–3 (your dataset and your supervised and unsupervised experiments). The paper should describe both what you did technically and what you learned from the relative performance of the machine learning approaches you applied to your dataset. This assignment should be posted as a PDF in your GitHub repository.

1) Describe your data set, why you chose it, and what you are trying to predict with it

2) Detail what you did to clean your data and any changes in the data representation that you applied. Discuss any challenges that arose during this process.

3) Discuss what you learned by visualizing your data

4) Describe your experiments with the two supervised learning algorithms you chose. This should include a brief description, in your own words, of what the algorithms do and the parameters that you adjusted. You should also report the relative performance of the algorithms on predicting your target attribute, reflecting on the reasons for any differences in performance between models and parameter settings.

Final Paper (Due July 8)

5) Describe your experiments using PCA for feature selection, discussing whether it improved any of your results with your best-performing supervised learning algorithm.

6) Discuss the results of using PCA as a pre-processing step for clustering. This should include a brief description, in your own words, of what the algorithms do. If you used the Wine dataset for this, briefly explain why your original dataset wasn't appropriate.

7) Summarize what you learned across the three projects, including what you think worked and what you would do differently if you had to do it over.

Final Paper Rubric (Due July 8)

	Missing	Fair	Good	Very Good	Excellent
Describe your data set, why you chose it, and what you are trying to predict with it.	**0** (0.00%)	**1.7** (8.50%) Something missing and cursory discussion	**1.8** (9.00%) Something missing or cursory d scussion	**1.9** (9.50%) Everything discussed in detai.	**2** (10.00%) Everything discussed in detail with nsight.
Detail what you did to clean your data and any changes in the data representation that you applied. Discuss any challenges that arose during this process.	**0** (0.00%)	**1.7** (8.50%) Something missing and cursory discussion	**1.8** (9.00%) Something missing or cursory d scussion	**1.9** (9.50%) Everything discussed in detai.	**2** (10.00%) Everything discussed in detail with nsight.
Discuss what you learned by visualizing your data	**0** (0.00%)	**1.7** (8.50%) Something missing and cursory discussion	**1.8** (9.00%) Something missing or cursory d scussion	**1.9** (9.50%) Everything discussed in detai.	**2** (10.00%) Everything discussed in detail with nsight.
Describe your experiments with the two supervised learning algorithms you chose. This should include a brief description, in your own words, of what the algorithms do and the parameters that you adjusted. You should also report the relative performance of the algorithms on predicting your target attribute, reflecting on the reasons for any differences in performance between models and parameter settings.	**0** (0.00%)	**3.4** (17.00%) Something missing and cursory discussion	**3.6** (18.00%) Something missing or cursory d scussion	**3.8** (19.00%) Everything discussed in detai.	**4** (20.00%) Everything discussed in detail with nsight.

Final Paper Rubric (Due July 8)

Describe your experiments using PCA for feature selection, discussing whether it improved any of your results with your best performing supervised learning algorithm.	**0** (0.00%)	**2.55** (12.75%) Something missing and cursory discussion	**2.7** (13.50%) Something missing or cursory discussion	**2.85** (14.25%) Everything discussed in detail.	**3** (15.00%) Everything discussed in detail with insight.
Discuss the results of using PCA as a pre-processing step for clustering. This should include a brief description, in your own words, of what the algorithms do. If you used the Wine dataset for this, briefly explain why your original dataset wasn't appropriate	**0** (0.00%)	**2.55** (12.75%) Something missing and cursory discussion	**2.7** (13.50%) Something missing or cursory discussion	**2.85** (14.25%) Everything discussed in detail.	**3** (15.00%) Everything discussed in detail with insight.
Summarize what you learned across the three projects, including what you think worked and what you would do differently if you had to do it over.	**0** (0.00%)	**1.7** (8.50%) Something missing and cursory discussion	**1.8** (9.00%) Something missing or cursory discussion	**1.9** (9.50%) Everything discussed in detail.	**2** (10.00%) Everything discussed in detail with insight.
Overall accuracy/depth of understanding and flow of the paper	**0** (0.00%)	**1.7** (8.50%) Numerous issues with the description of contents/flow of the paper	**1.8** (9.00%) Some issues with the description of contents/flow of the paper	**1.9** (9.50%) Accurately describe all the relevant concepts/flow of the paper	**2** (10.00%) Excellent discussion of relevant concepts/flow of the paper