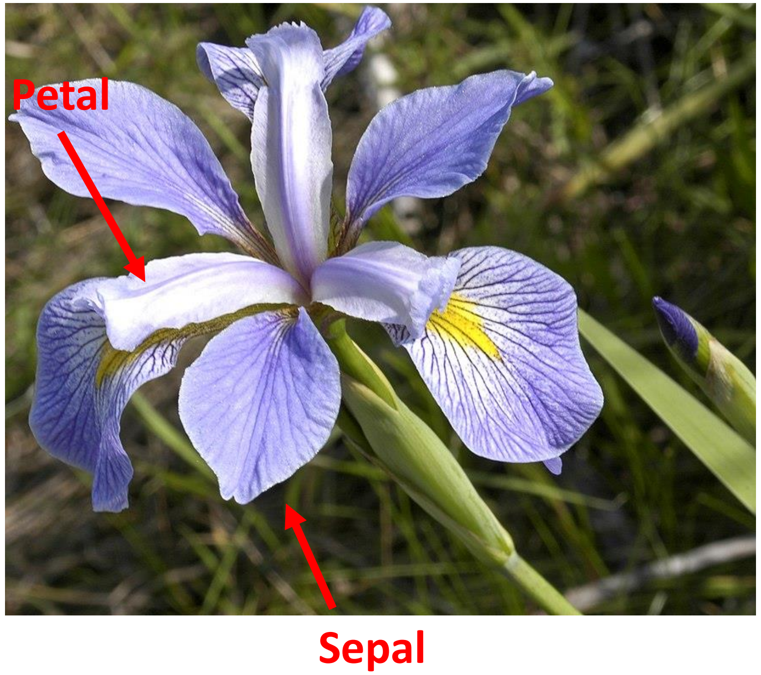
Section 10: Case Studies

Section 10.1: Clustering for Iris Dataset

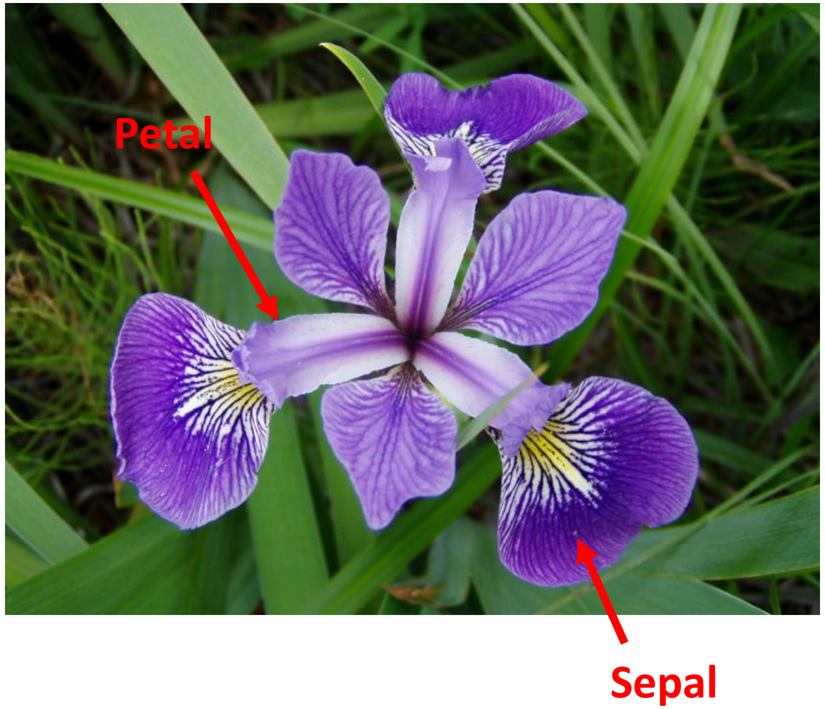
Iris Flower Dataset

Three types of iris flowers in dataset

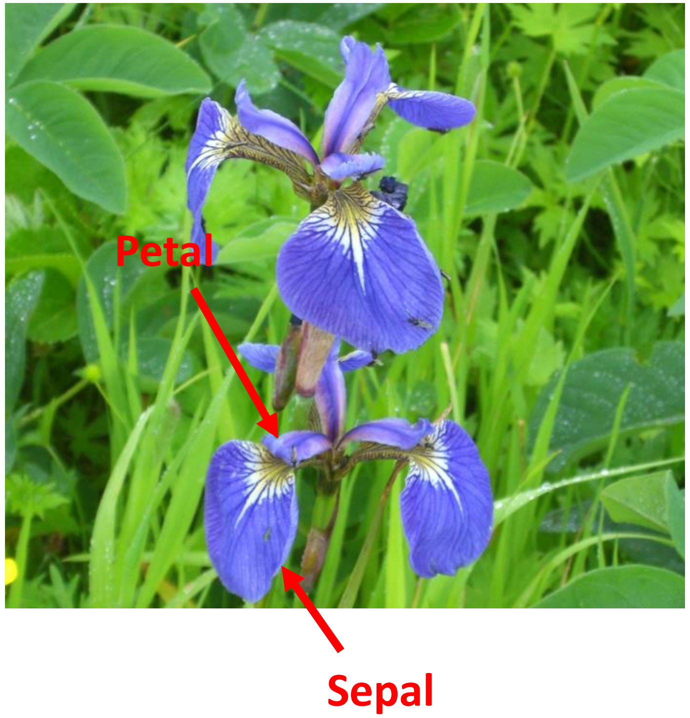
Iris Virginica



Iris Versicolor



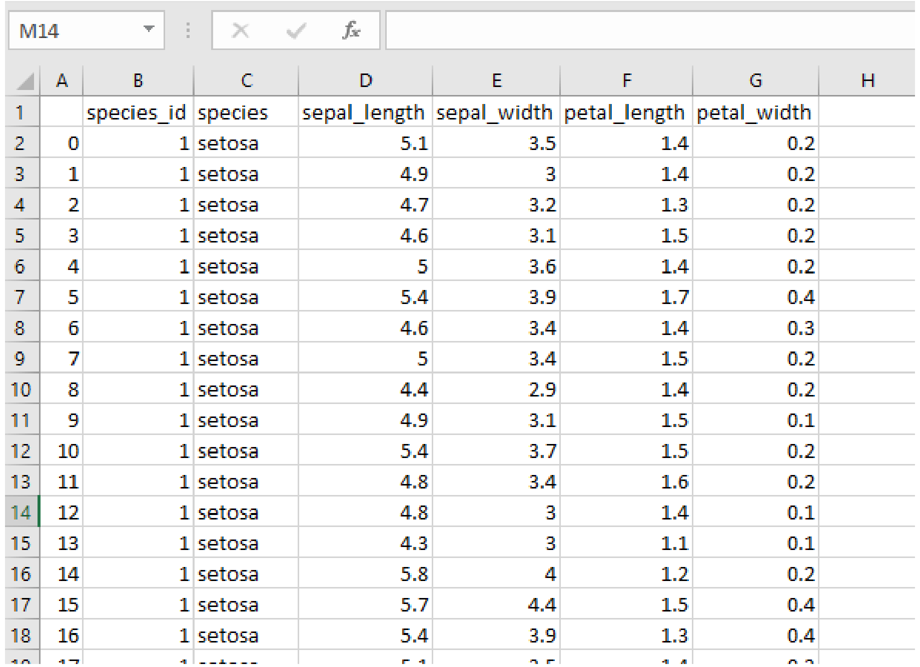
Iris Setosa



See UnsupervisedML_Resources.pdf file for links
Images reproduced here under Wikipedia Commons Copyright

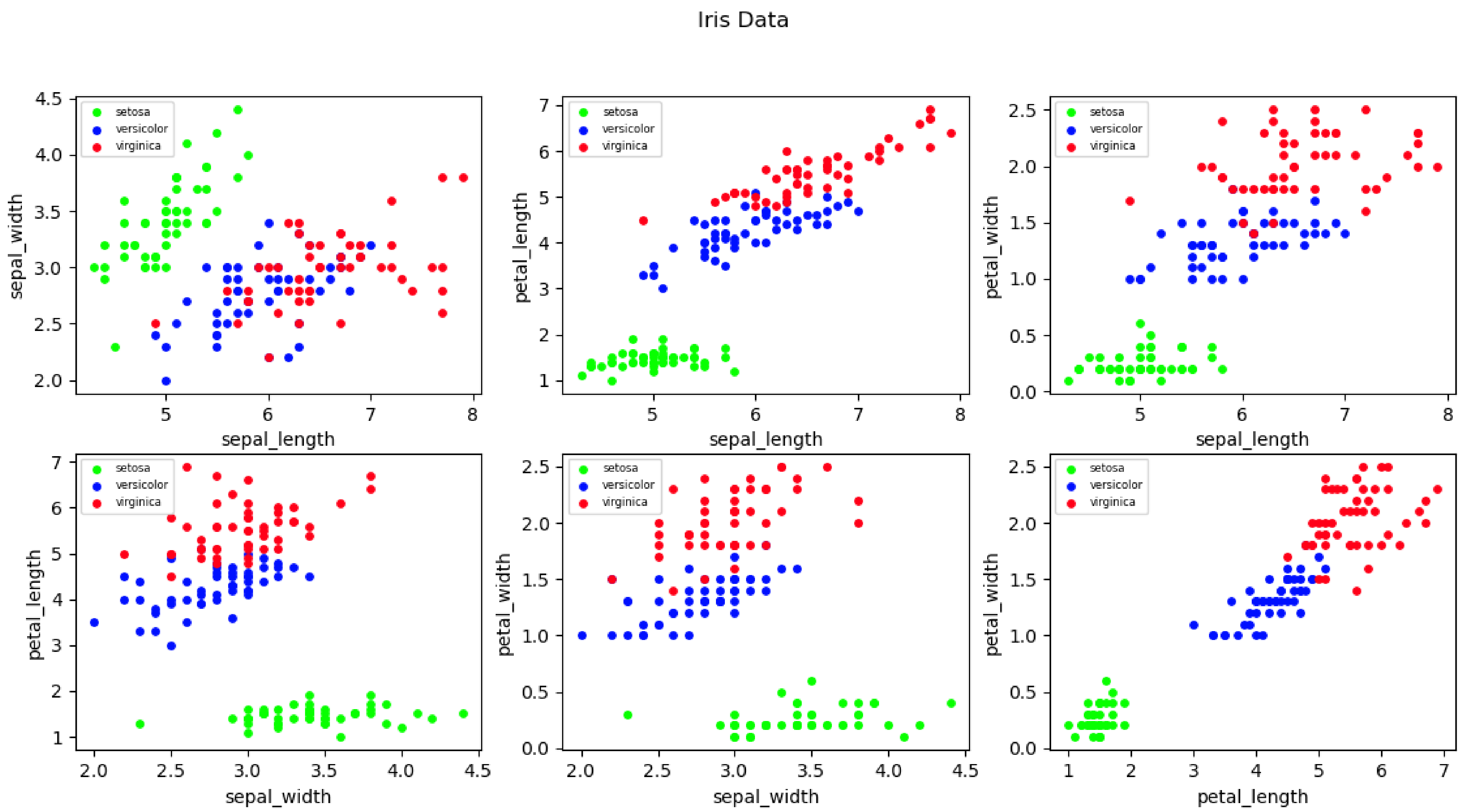
Iris Dataset

- 50 samples each of 3 types of iris flower species: setosa, virginica, versicolor
- 4 features: sepal_length, sepal_width, petal_length, petal_width
- Species id and species columns give labels (typically used in Supervise Learning)
- Dataset available at UCI, Irvine, Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/iris
- File: Unsupervised/Clustering/Code/Data_Iris/Iris.csv



	A	B	C	D	E	F	G	H
1		species_id	species	sepal_length	sepal_width	petal_length	petal_width	
2	0	1	setosa	5.1	3.5	1.4	0.2	
3	1	1	setosa	4.9	3	1.4	0.2	
4	2	1	setosa	4.7	3.2	1.3	0.2	
5	3	1	setosa	4.6	3.1	1.5	0.2	
6	4	1	setosa	5	3.6	1.4	0.2	
7	5	1	setosa	5.4	3.9	1.7	0.4	
8	6	1	setosa	4.6	3.4	1.4	0.3	
9	7	1	setosa	5	3.4	1.5	0.2	
10	8	1	setosa	4.4	2.9	1.4	0.2	
11	9	1	setosa	4.9	3.1	1.5	0.1	
12	10	1	setosa	5.4	3.7	1.5	0.2	
13	11	1	setosa	4.8	3.4	1.6	0.2	
14	12	1	setosa	4.8	3	1.4	0.1	
15	13	1	setosa	4.3	3	1.1	0.1	
16	14	1	setosa	5.8	4	1.2	0.2	
17	15	1	setosa	5.7	4.4	1.5	0.4	
18	16	1	setosa	5.4	3.9	1.3	0.4	

Iris Dataset



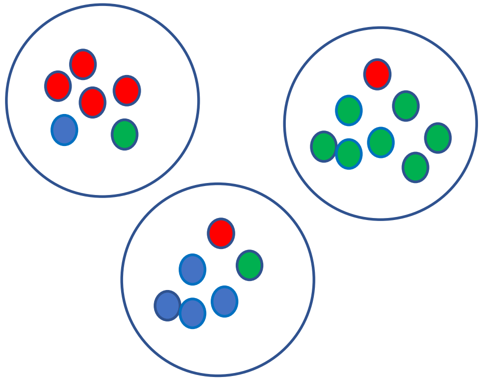
Metrics for Measuring Quality

- Purity measures extent to which clusters contain a single class
- Useful for testing purposes if class labels are provided
  - M is number of data points, C is set of clusters, D is set of classes
  - For each cluster: determine maximum number of data points from any one class and sum over all clusters

$$P = \frac{1}{M} \sum_{c \in C} max_{d \in D} |d \cap c|$$

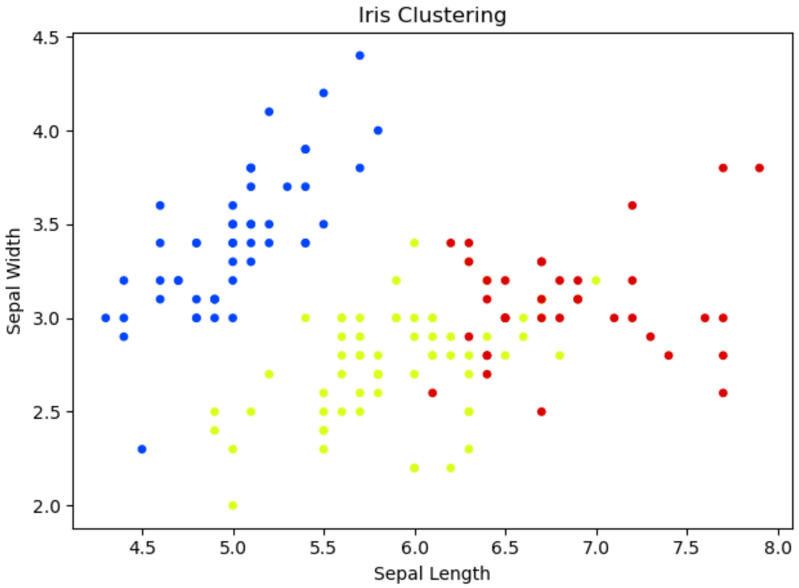
  - Purity satisfies $0 < P \leq 1$

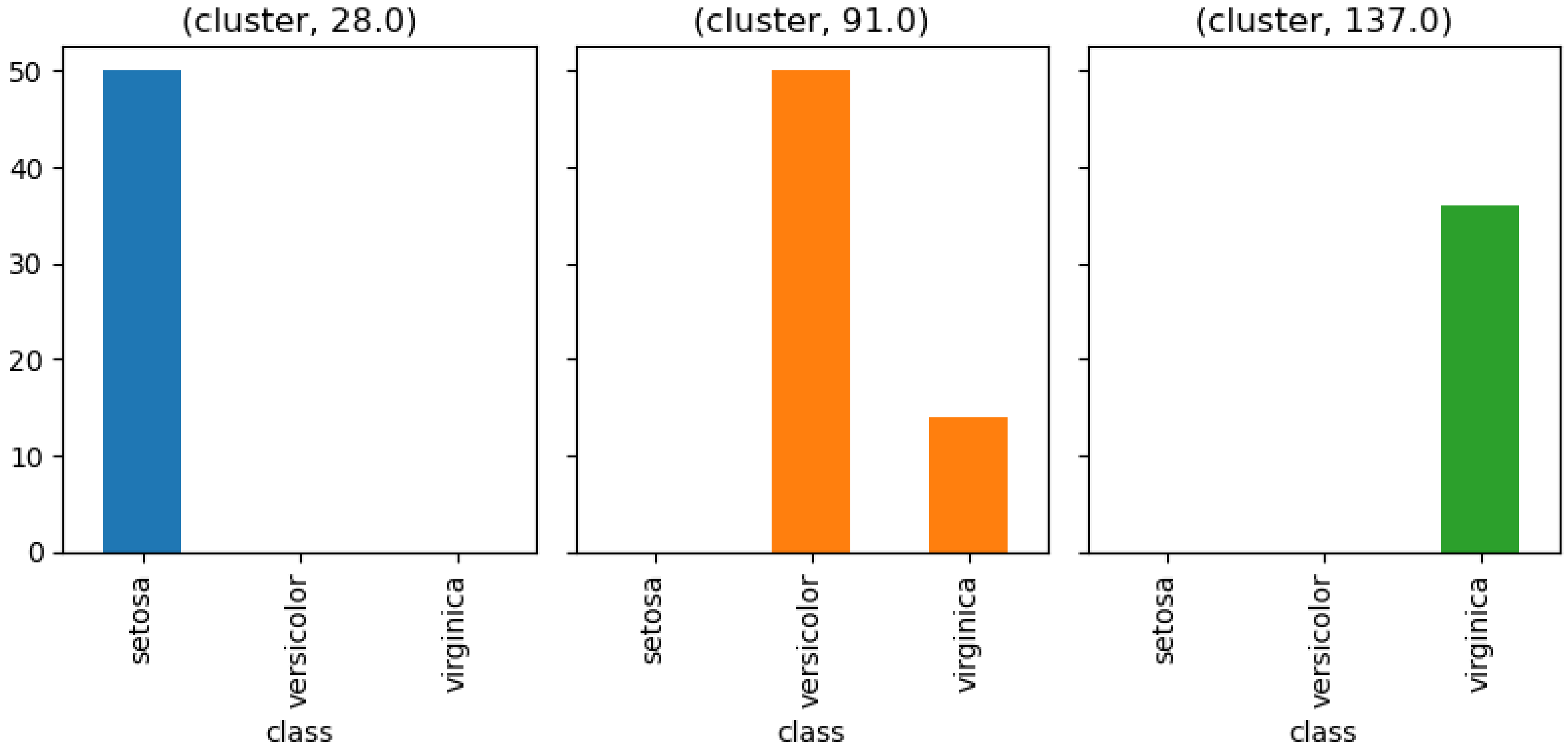
- Example
  - 20 data points and 3 clusters
  - 3 Actual Classes: red, blue, green
  - Max from any class:
    - Cluster 1: 4 red, Cluster2: 4 blue, Cluster 3: 7 green
  - $P = \frac{1}{20}(4 + 4 + 7) = 0.75$



Clustering for Iris Dataset

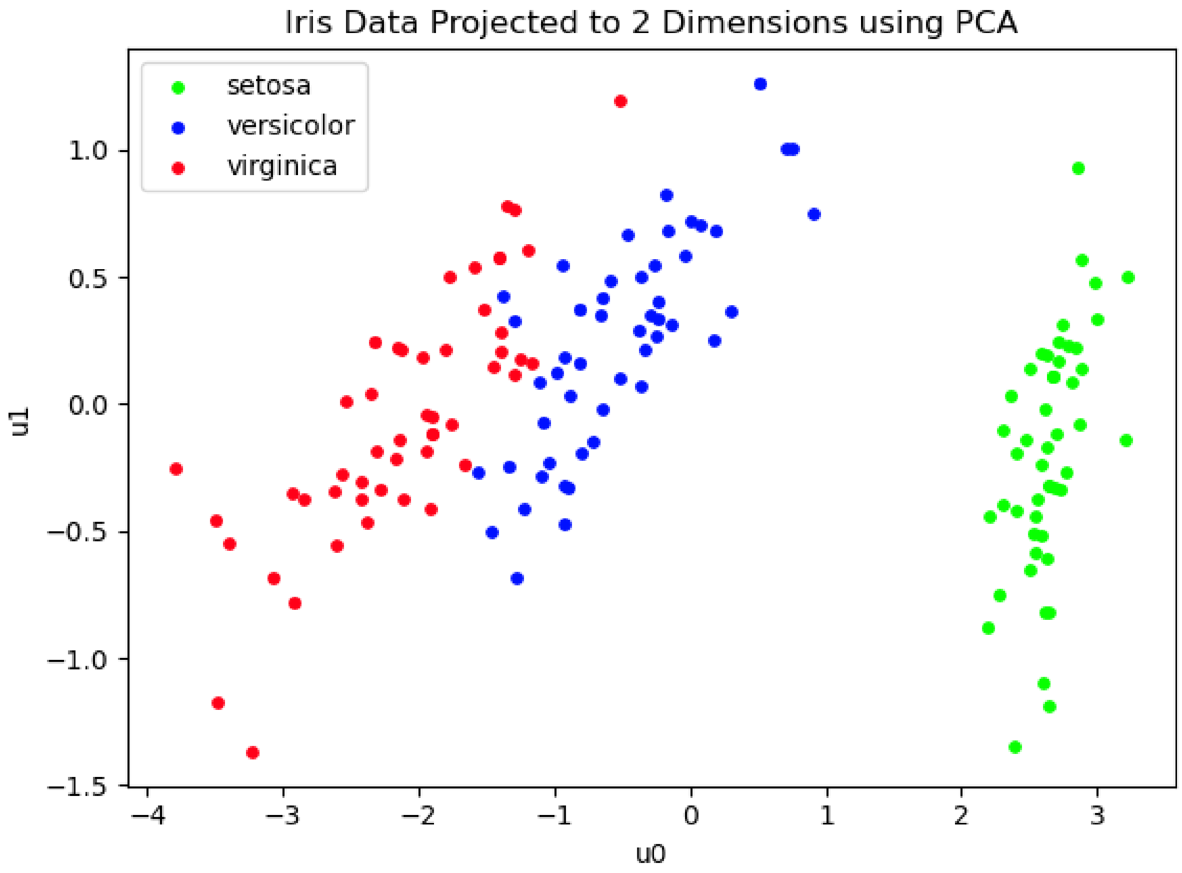
- Algorithm: Hierarchical stopping at 3 clusters
- Metrics:
  - Purity: 0.907
  - Silhouette: 0.554





PCA for Iris Dataset

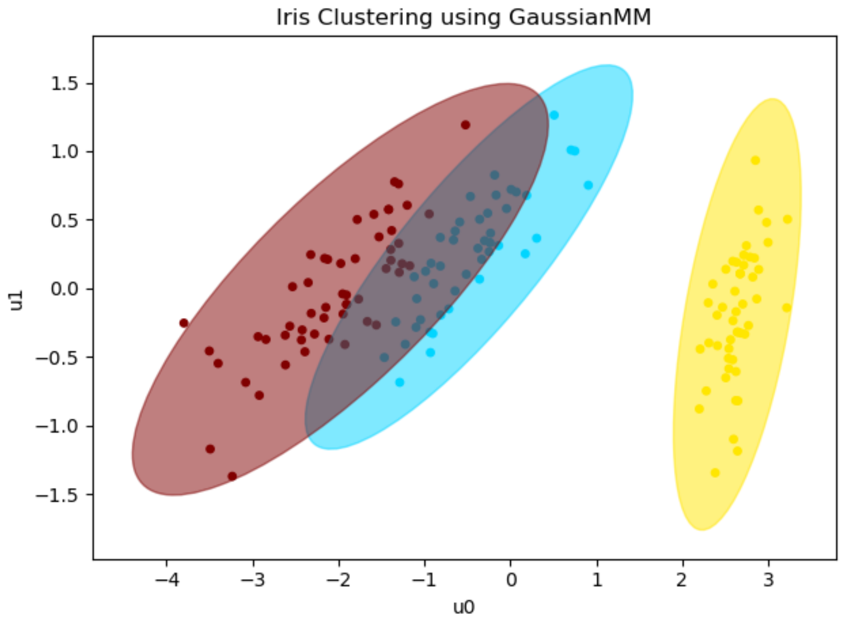
- Project data from 4 dimensions to 2 dimension using PCA
- Variance capture is 97.8%

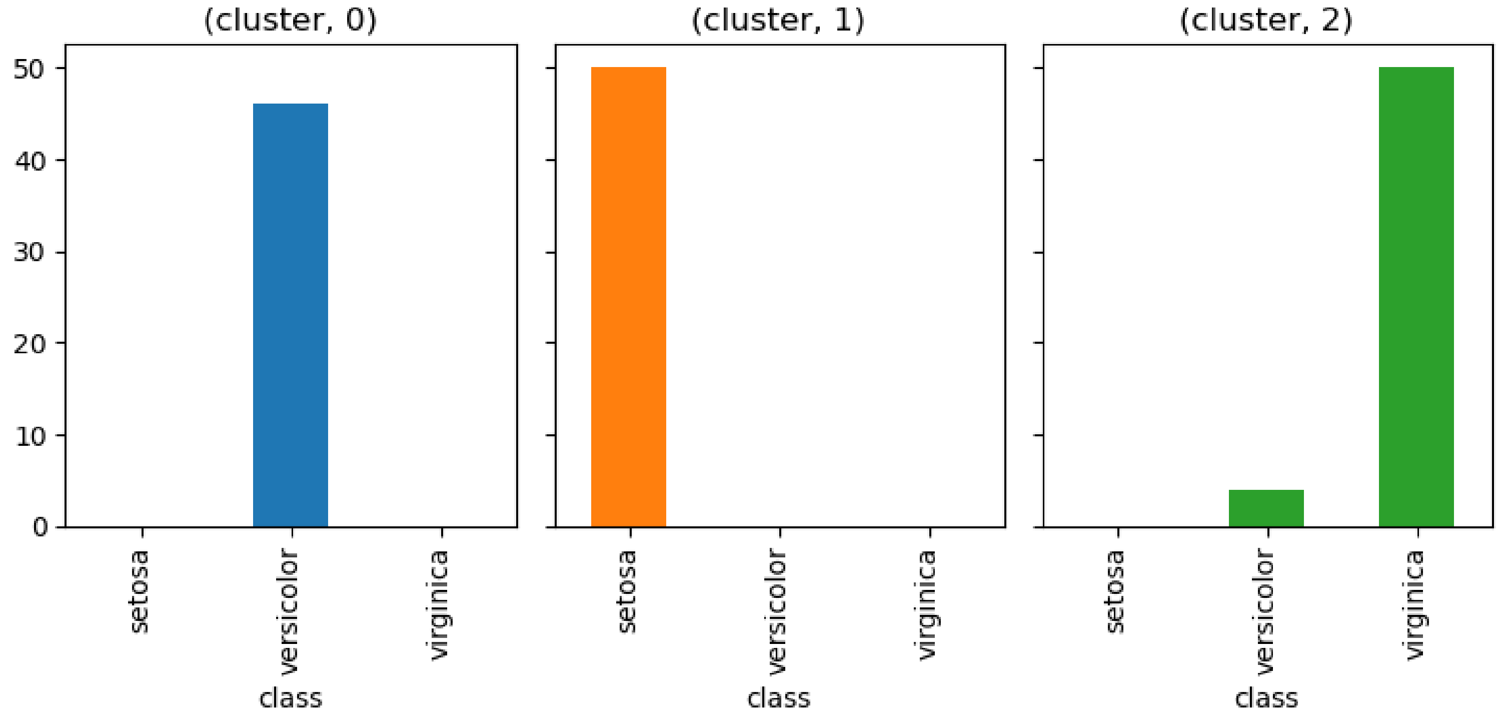


- New features u0 and u1 do not correspond to actual measurable quantity, such as sepal width/length or petal width/length

Clustering for Iris Dataset

- Algorithm: Gaussian Mixture Model
  - Specify 3 clusters and use K Means++ for initialization
- Metrics:
  - Purity: 0.973
  - Silhouette: 0.537

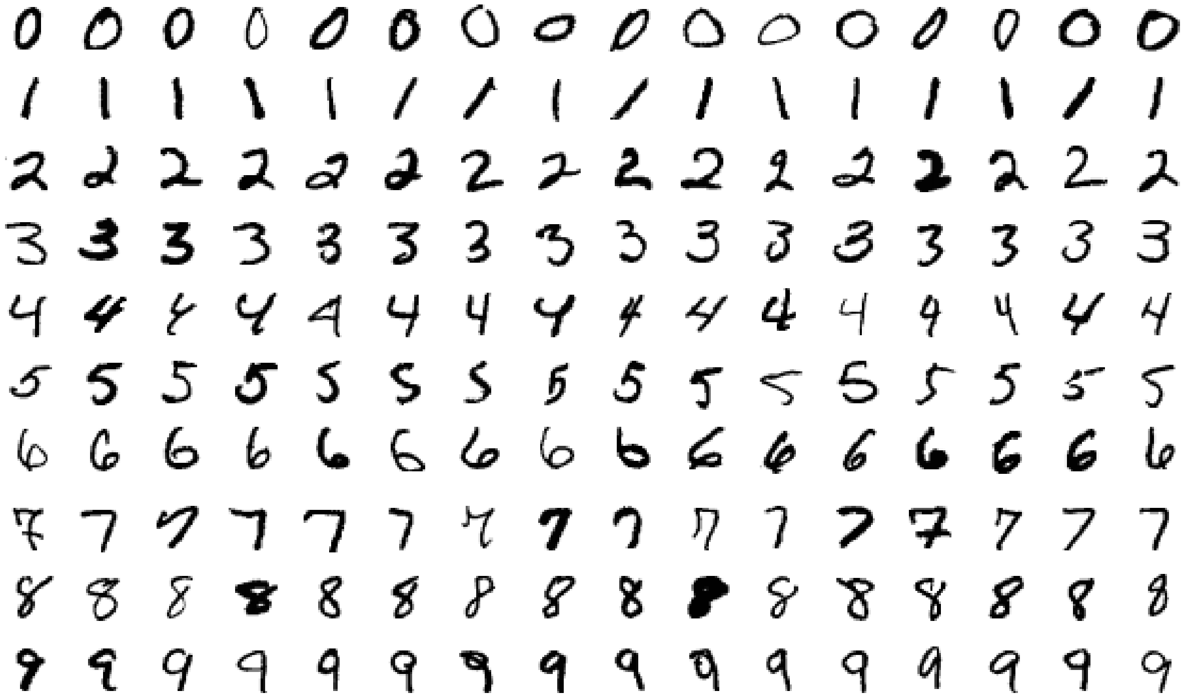




Section 10.2: Clustering for MNIST Digits Dataset

MNIST Digits Dataset

- Thousands of handwritten digit images with 28x28 resolution
- Data Source: http://yann.lecun.com/exdb/mnist/
- Used extensively for testing machine learning algorithms

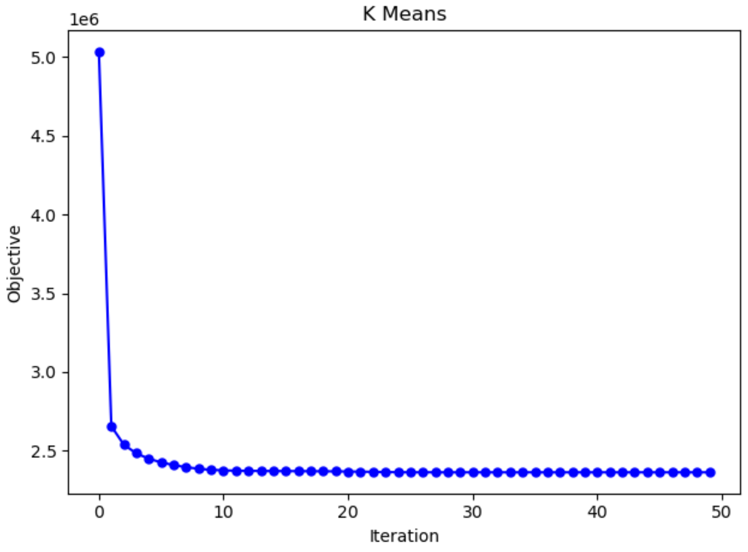


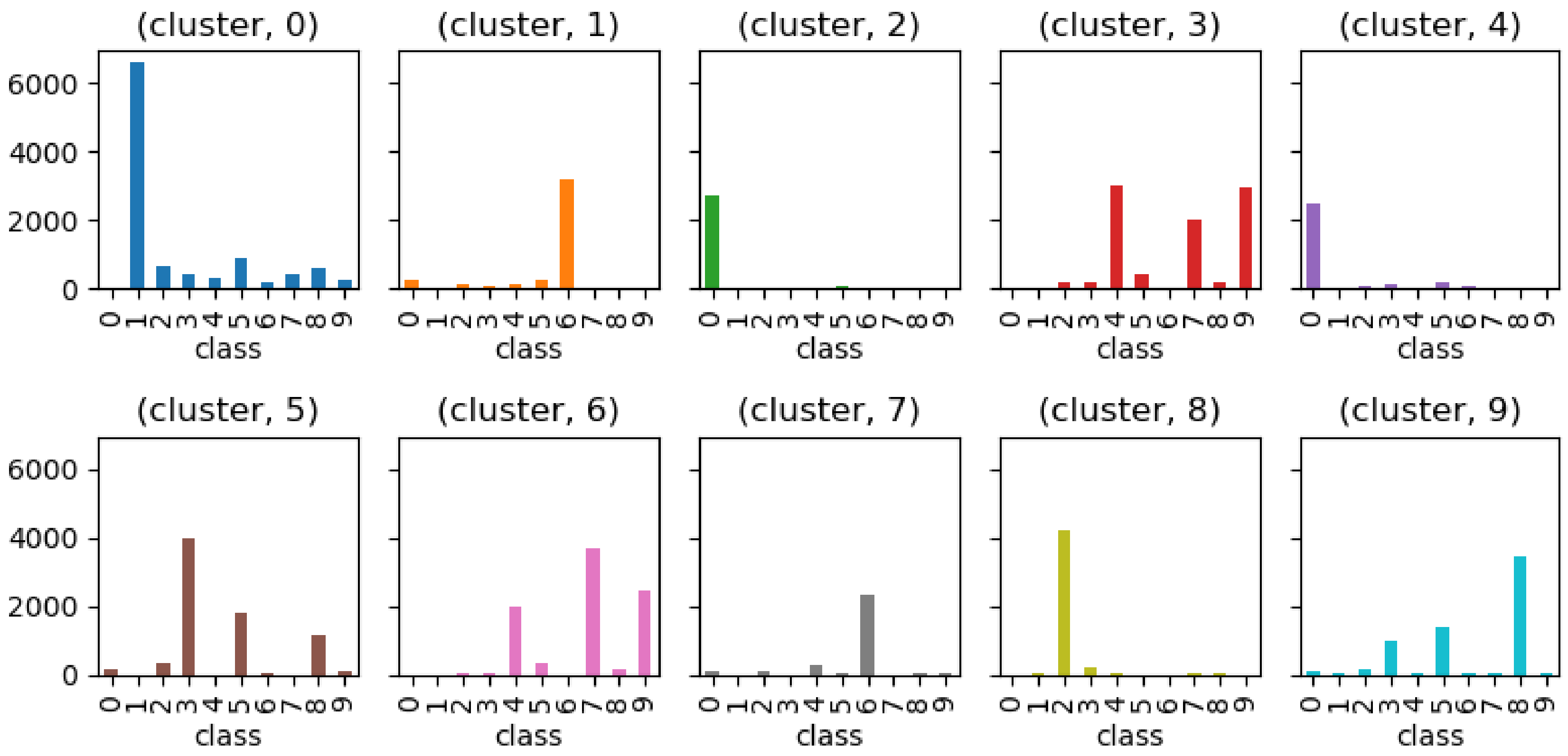
Collage of 160 individual digit images

By Josef Steppan - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=64810040

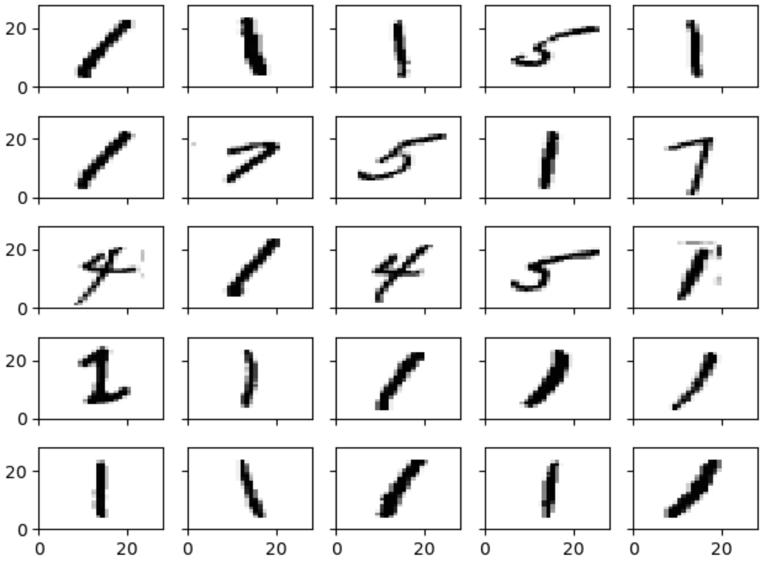
K Means Clustering for MNIST Digits

- 60000 images
- Algorithm: K Means with 10 clusters and K Means ++ for initialization
- Metrics:
    - Purity: 0.596
    - Run Time: 319 seconds

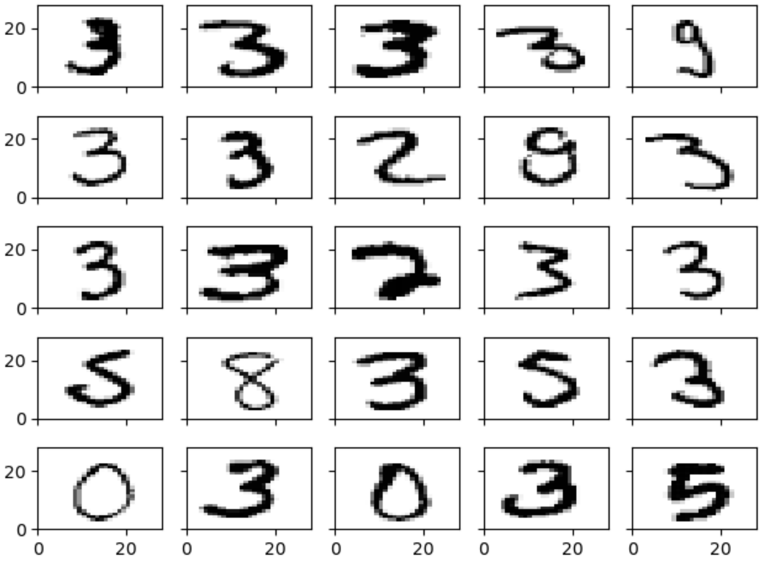




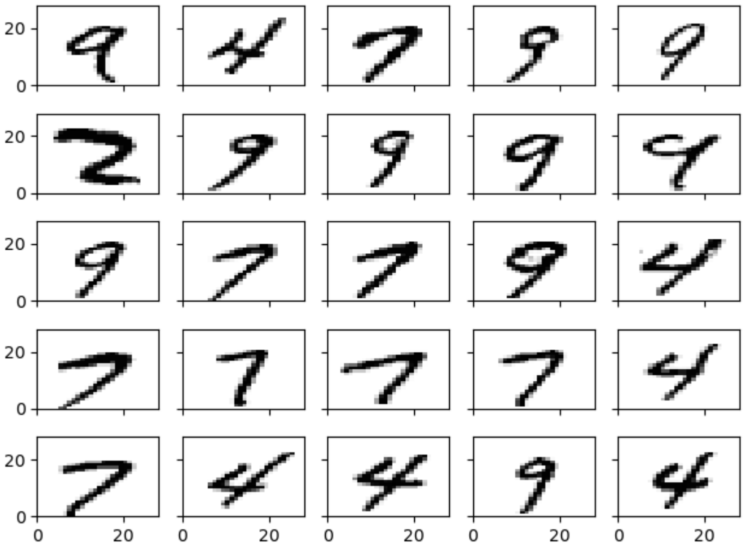
K MNIST Digits Clustering Results



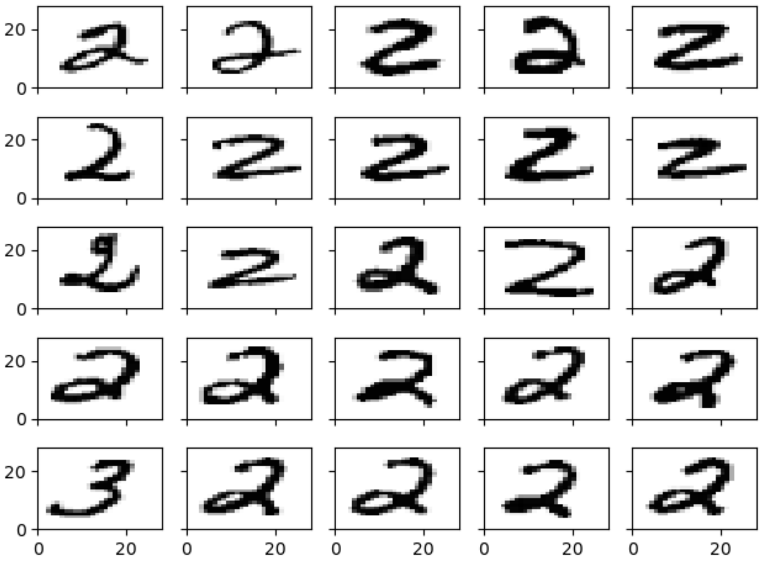
Cluster 0



Cluster 5



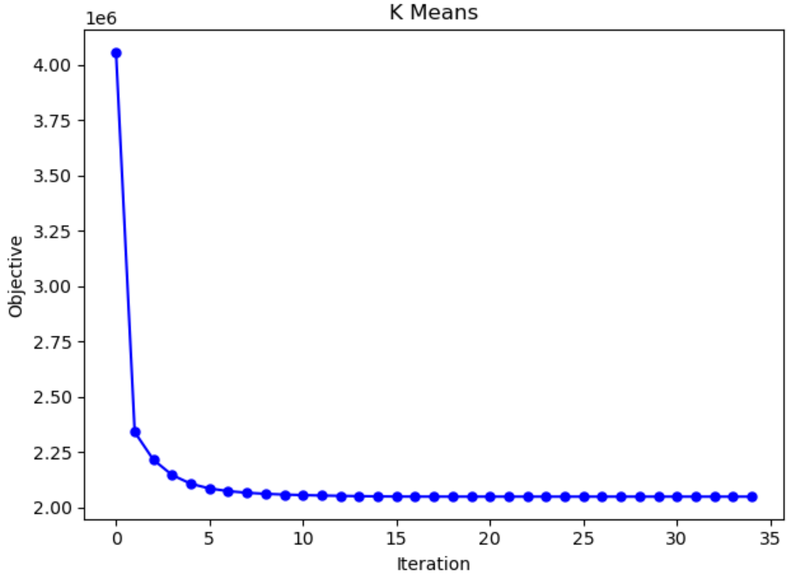
Cluster 6

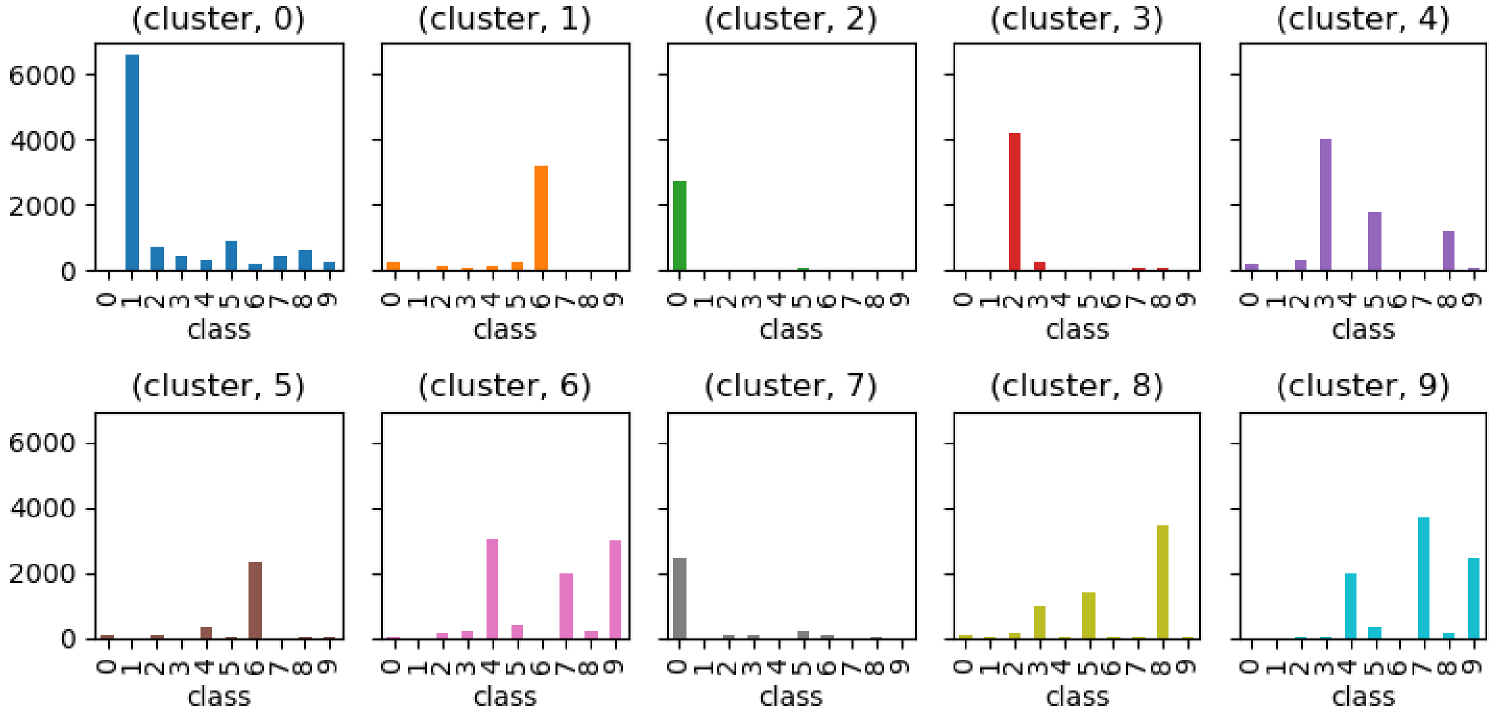


Cluster 8

K Means Clustering for MNIST Digits with PCA

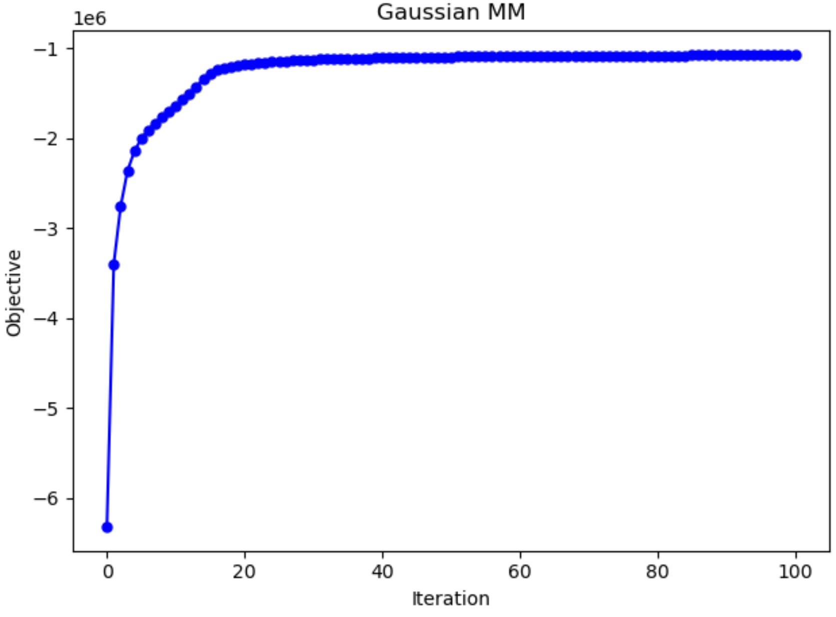
- Apply PCA with 90% variance capture (reduced from 784 to 87 dimensions)
- Algorithm: K Means with 10 clusters and K Means ++ for initialization
- Metrics:
  - Purity: 0.596
  - Run Time: 27 seconds

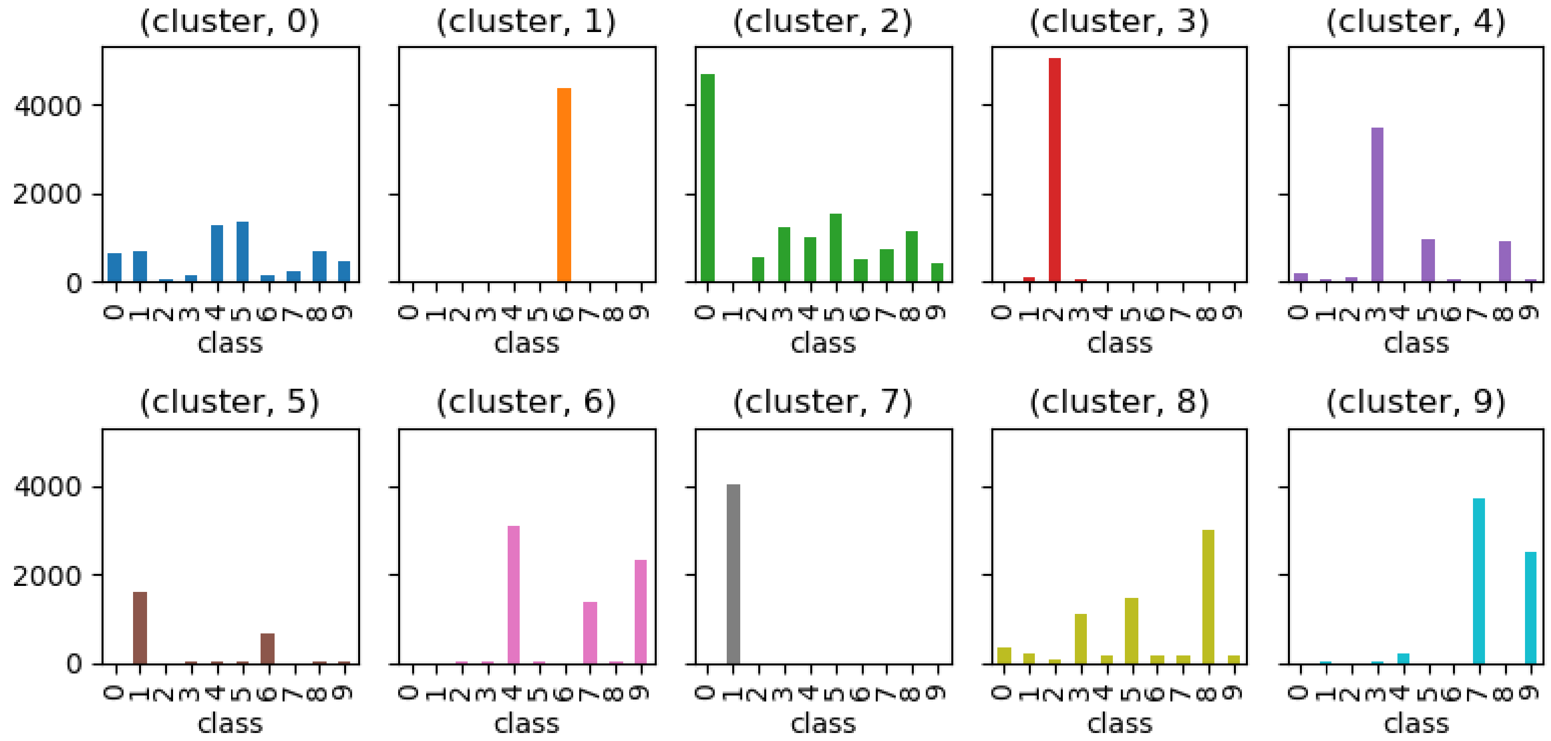




GMM Clustering for MNIST Digits with PCA

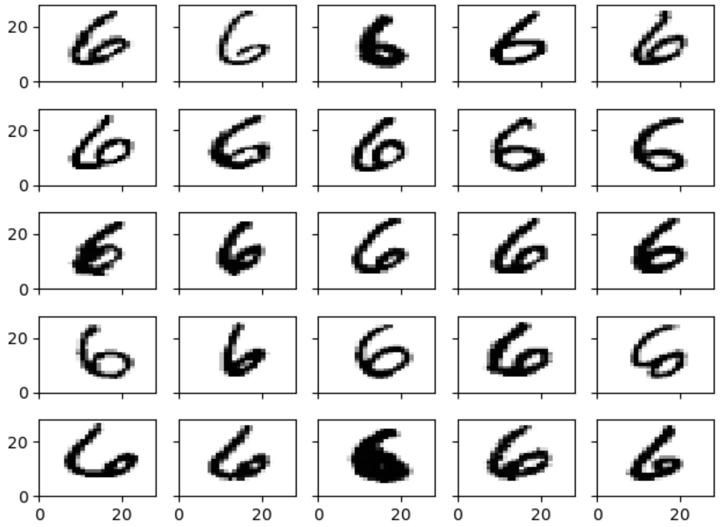
- Apply PCA with 90% variance capture (reduced from 784 to 87 dimensions)
- Algorithm: Gaussian MM with 10 clusters and K Means ++ for initialization
- Metrics:
  - Purity: 0.574
  - Run Time: 264 seconds



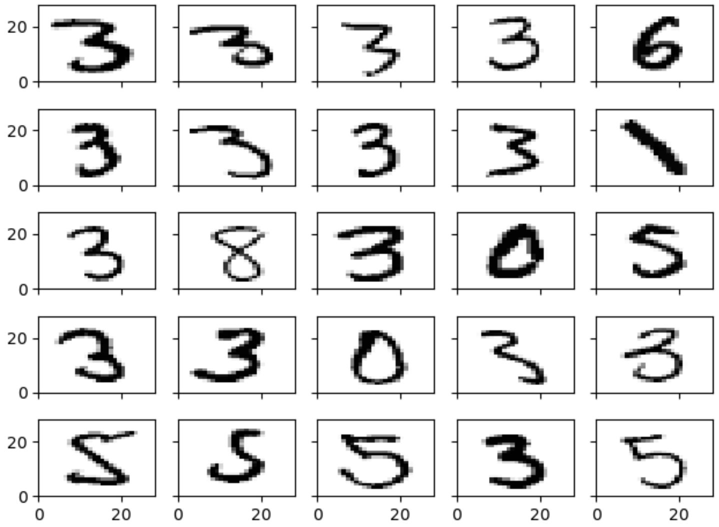


MNIST Clustering Results: Clusters

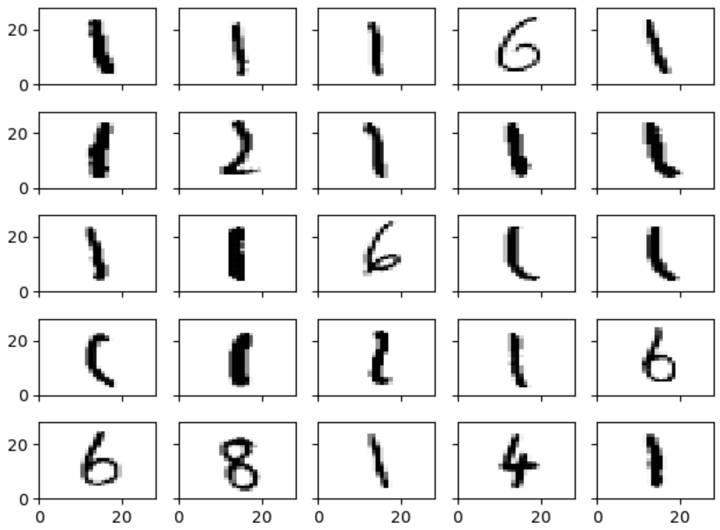
Cluster 1



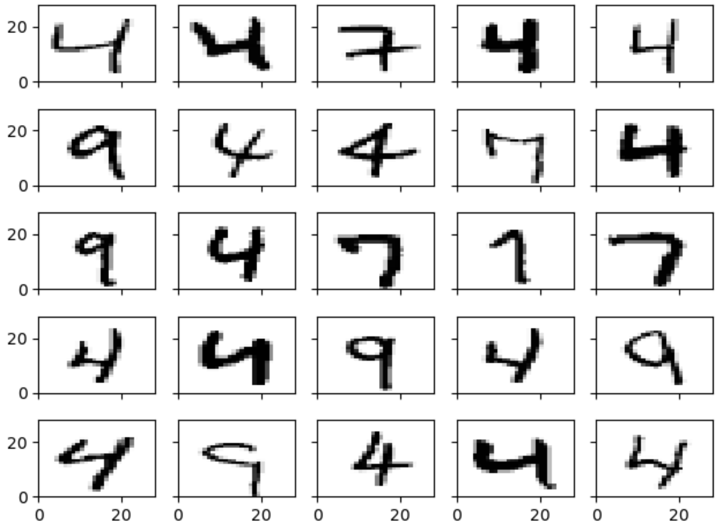
Cluster 4



Cluster 5



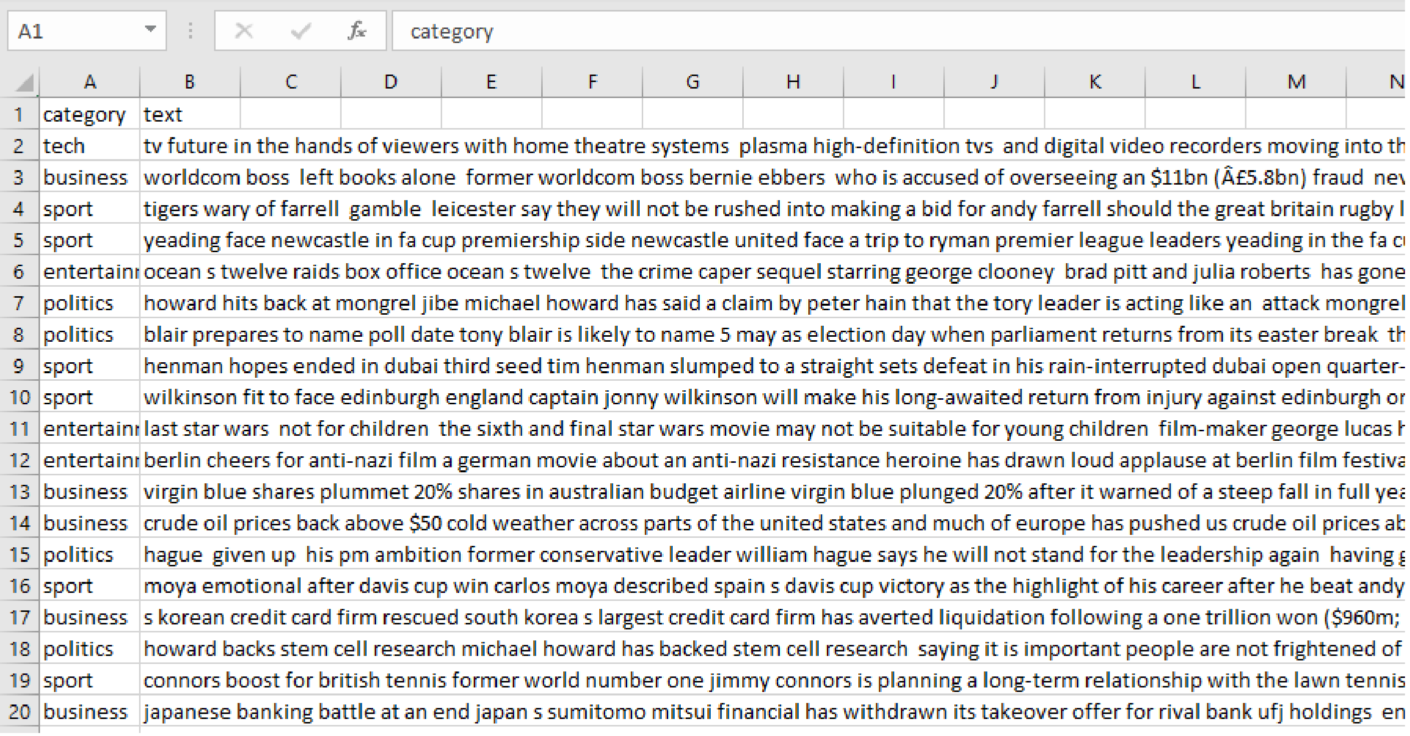
Cluster 6



Section 10.3: Clustering for Text Documents

BBC News Text Dataset

- 2225 news articles
- 5 classes: sports, business, tech, entertainment, politics
- Dataset from Kaggle
- https://www.kaggle.com/yufengdev/bbc-fulltext-and-category
- File: Unsupervised/Clustering/Data_Text/bbc-text.csv
- Use Tfidf vectorizer in sklearn
- 12915 words in dictionary
- 12915 x 2225 feature matrix



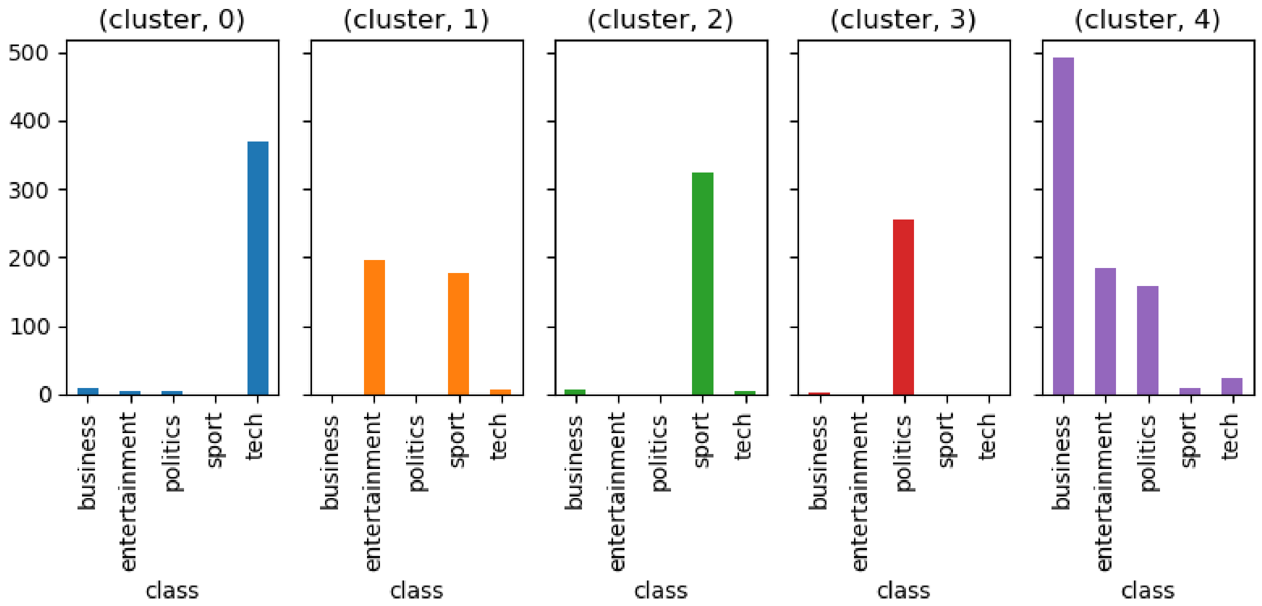
A1 category

	A	B
1	category	text
2	tech	tv future in the hands of viewers with home theatre systems plasma high-definition tvs and digital video recorders moving into th
3	business	worldcom boss left books alone former worldcom boss bernie ebbers who is accused of overseeing an $11bn (Â£5.8bn) fraud nev
4	sport	tigers wary of farrell gamble leicester say they will not be rushed into making a bid for andy farrell should the great britain rugby l
5	sport	yeading face newcastle in fa cup premiership side newcastle united face a trip to ryman premier league leaders yeading in the fa c
6	entertainı	ocean s twelve raids box office ocean s twelve the crime caper sequel starring george clooney brad pitt and julia roberts has gone
7	politics	howard hits back at mongrel jibe michael howard has said a claim by peter hain that the tory leader is acting like an attack mongrel
8	politics	blair prepares to name poll date tony blair is likely to name 5 may as election day when parliament returns from its easter break th
9	sport	henman hopes ended in dubai third seed tim henman slumped to a straight sets defeat in his rain-interrupted dubai open quarter-
10	sport	wilkinson fit to face edinburgh england captain jonny wilkinson will make his long-awaited return from injury against edinburgh or
11	entertainı	last star wars not for children the sixth and final star wars movie may not be suitable for young children film-maker george lucas h
12	entertainı	berlin cheers for anti-nazi film a german movie about an anti-nazi resistance heroine has drawn loud applause at berlin film festiva
13	business	virgin blue shares plummet 20% shares in australian budget airline virgin blue plunged 20% after it warned of a steep fall in full yea
14	business	crude oil prices back above $50 cold weather across parts of the united states and much of europe has pushed us crude oil prices ab
15	politics	hague given up his pm ambition former conservative leader william hague says he will not stand for the leadership again having g
16	sport	moya emotional after davis cup win carlos moya described spain s davis cup victory as the highlight of his career after he beat andy
17	business	s korean credit card firm rescued south korea s largest credit card firm has averted liquidation following a one trillion won ($960m;
18	politics	howard backs stem cell research michael howard has backed stem cell research saying it is important people are not frightened of
19	sport	connors boost for british tennis former world number one jimmy connors is planning a long-term relationship with the lawn tennis
20	business	japanese banking battle at an end japan s sumitomo mitsui financial has withdrawn its takeover offer for rival bank ufj holdings en

K Means Clustering for BBC News Text with PCA

- Algorithm: K Means with 5 clusters and K Means ++ for initialization
- Metrics:
  - Purity: 0.736
  - Fit Time: 42 seconds
- Record most influential words for each cluster

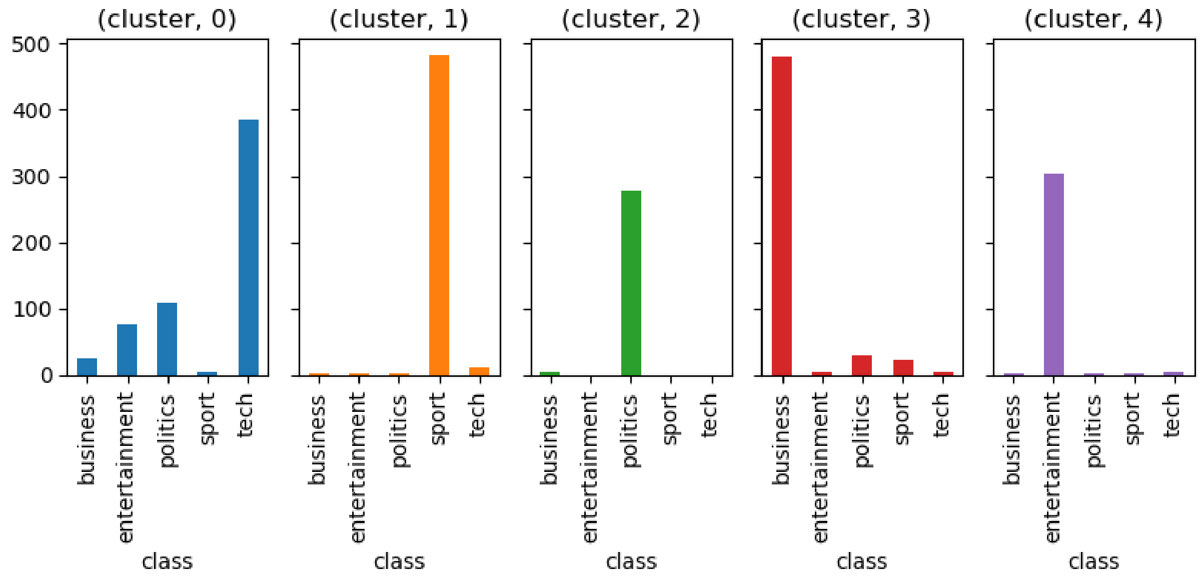
Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
people	film	england	mr	year
mobile	best	game	labour	mr
technology	year	club	election	government
users	awards	wales	blair	growth
software	won	chelsea	party	company
digital	award	rugby	brown	sales
music	world	players	howard	economy
net	champion	cup	government	new
games	festival	ireland	minister	bank
phone	films	team	tory	market



K Means Clustering for BBC News Text

- Apply PCA with 95% variance capture (reduced from 12915 to 1601 dimensions)
- Algorithm: K Means with 5 clusters and K Means ++ for initialization
- Metrics:
  - Purity: 0.867
  - Fit Time: 6 seconds
- Record most influential words for each cluster

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
people mobile music technology mr software users digital new games	game england win cup match team players injury play Club	mr labour election blair party brown government howard minister tory	growth economy year bank company market mr sales oil shares	Film best awards award band festival actor star album year


(cluster, 0)
(cluster, 1)
(cluster, 2)
(cluster, 3)
(cluster, 4)
business
entertainment
politics
sport
tech
class