ClusterCat Algorithm: Supervised Subcategory K-Means Clustering

Paul DiStefano
Binghamton University--SUNY

LaRC Lab
Learning and Representation in Cognition

ClusterCat Algorithm: Supervised Subcategory K-Means Clustering

Paul V DiStefano III and Kenneth J Kurtz
Department of Psychology, Binghamton University (SUNY)

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

Background

- Supervised learning uses labeled data to make classification decisions, while unsupervised learning uses unlabeled data
- K-means clustering is an unsupervised clustering algorithm that partitions data into k number of clusters
- Not all data is labelled and sometimes labels do not capture structure within categories

Motivation

- Subcategories within a label may provide useful information for generalizing knowledge to classify new points
- ClusterCat aims to utilize supervised learning to create subcategories, then cluster them using unsupervised learning

Methods

The ClusterCat Algorithm:

- First ,the dataset is split into training data (80%) and test data (20%). The training set is the partitioned by known category label. K-Means is performed on each category to create subcategories within each label
- The number of clusters is determined automatically using the silhouette score (a value that measures similarity of an object to its own cluster compared with other clusters).
- After ClusterCat completes the training phase, test points are classified into the created subcategories based on the following:

1.) If the test point is contained within the range of a subcategory, it is assigned there
2.) If the test point is contained within more than one subcategory range, it is assigned to the subcategory with the nearest prototype
3.) If the test point is not within any subcategory range, it is assigned to the subcategory with the nearest point

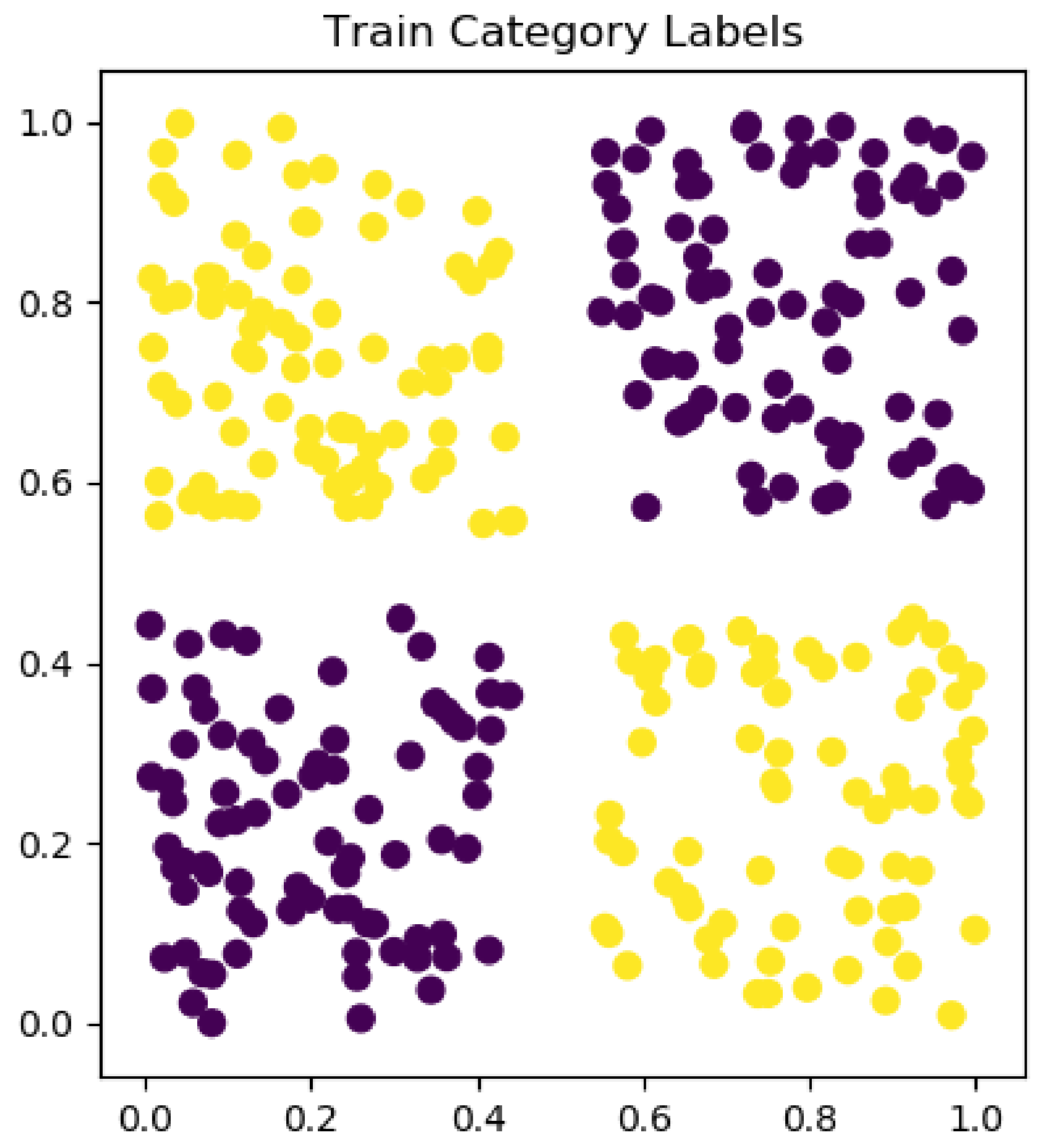
In summation, ClusterCat uses the subcategories to classify unseen items

Results

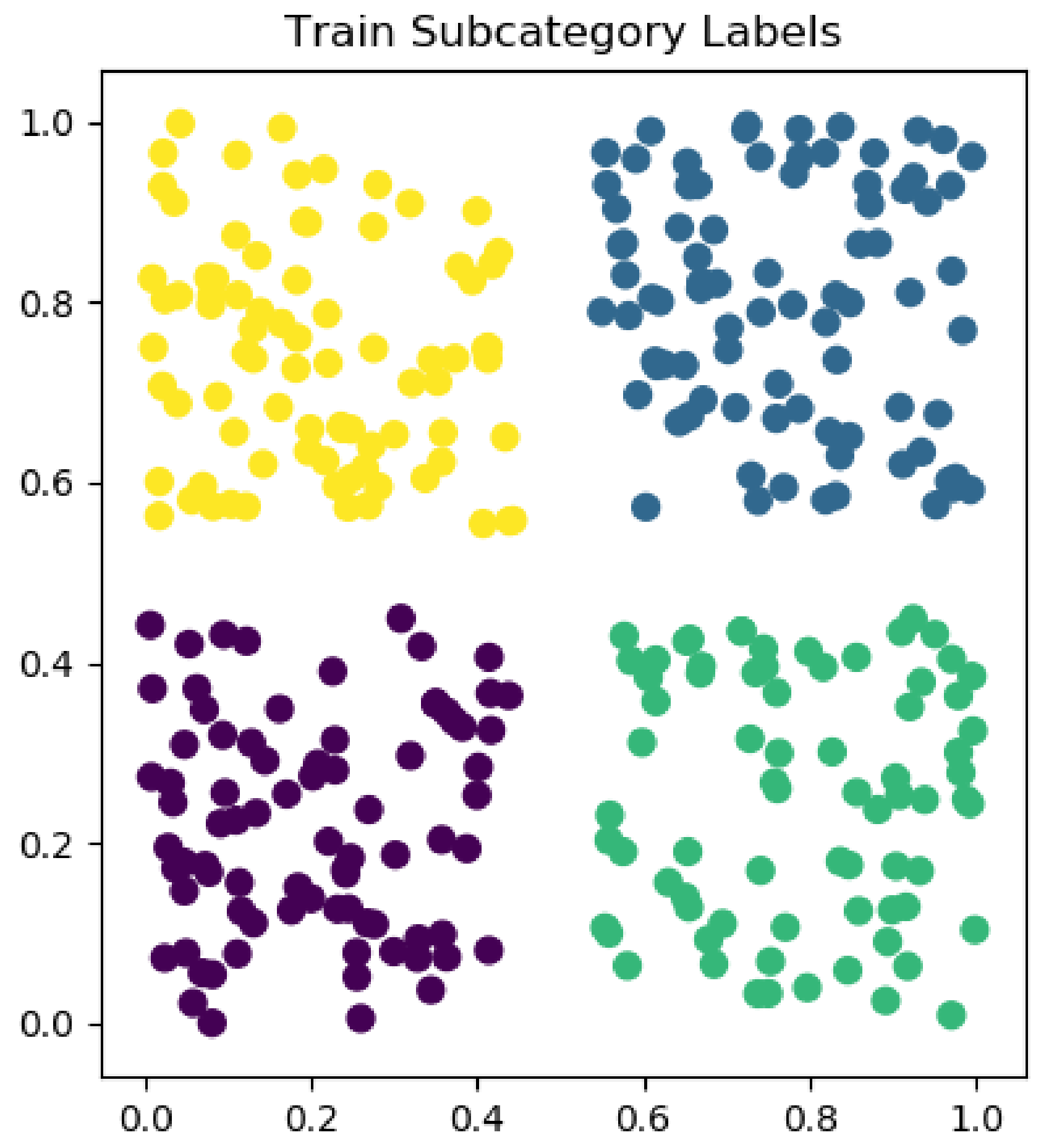
- The graphs to the right depict ClusterCat training on and then classifying an XOR dataset
- XOR, short for 'exclusive or', is a logic gate that reports true if the two conditions differ. ClusterCat creates subcategories within the XOR Dataset, which leads to near perfect accuracy
- Since ClusterCat can differentiate between clusters within the same category, it can accurately classify the XOR dataset while K-Means alone cannot
- Accuracies reported in the table below are the average accuracy score of ClusterCat on different UCI Datasets over 30 initializations with standard deviations shown
- ClusterCat has demonstrated reasonable accuracy with low variability between iterations
- More complex datasets (more features and non-linear separability) appear to have lower accuracy

XOR Dataset

Training Phase:

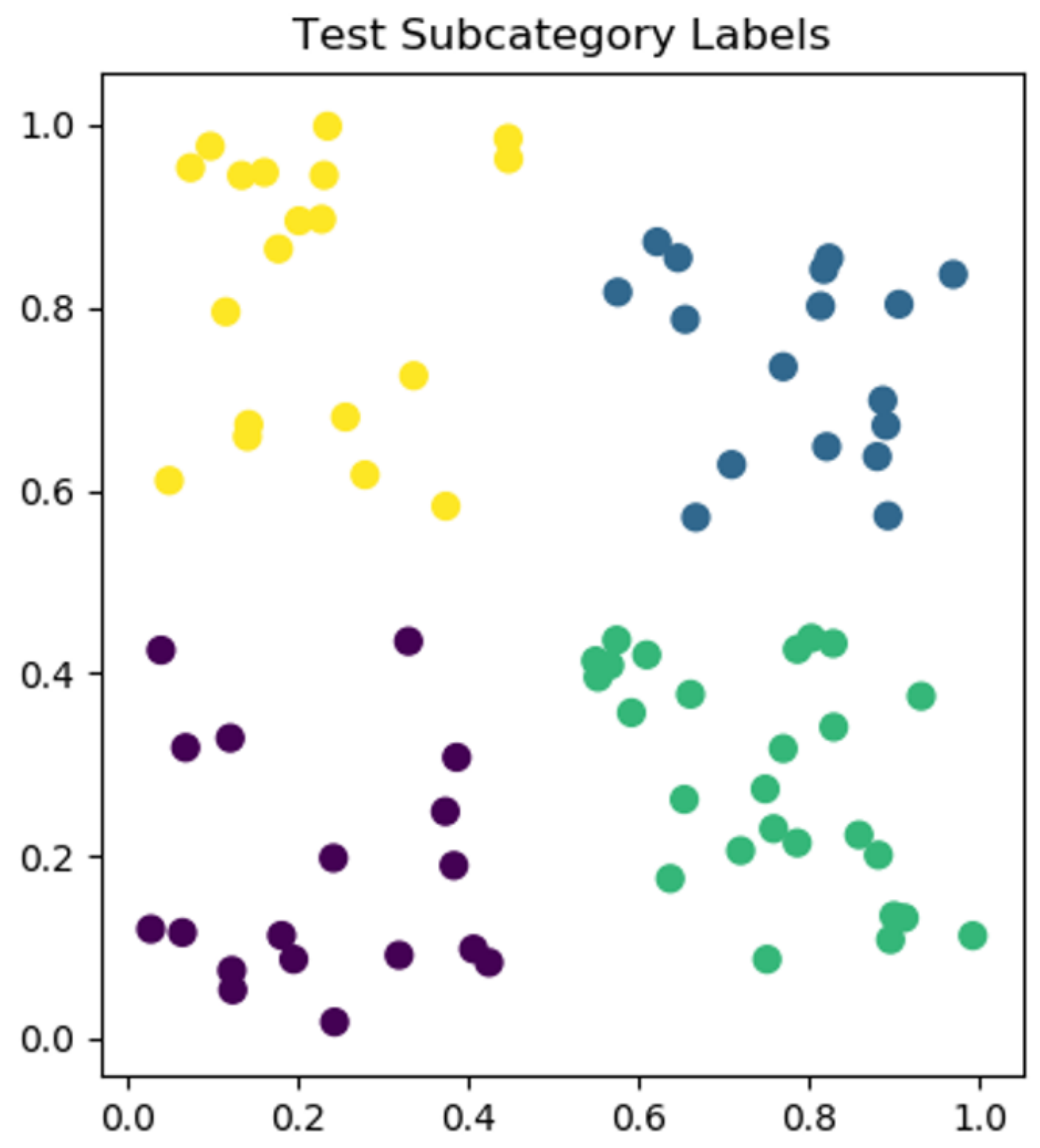


Dataset is split by known category label

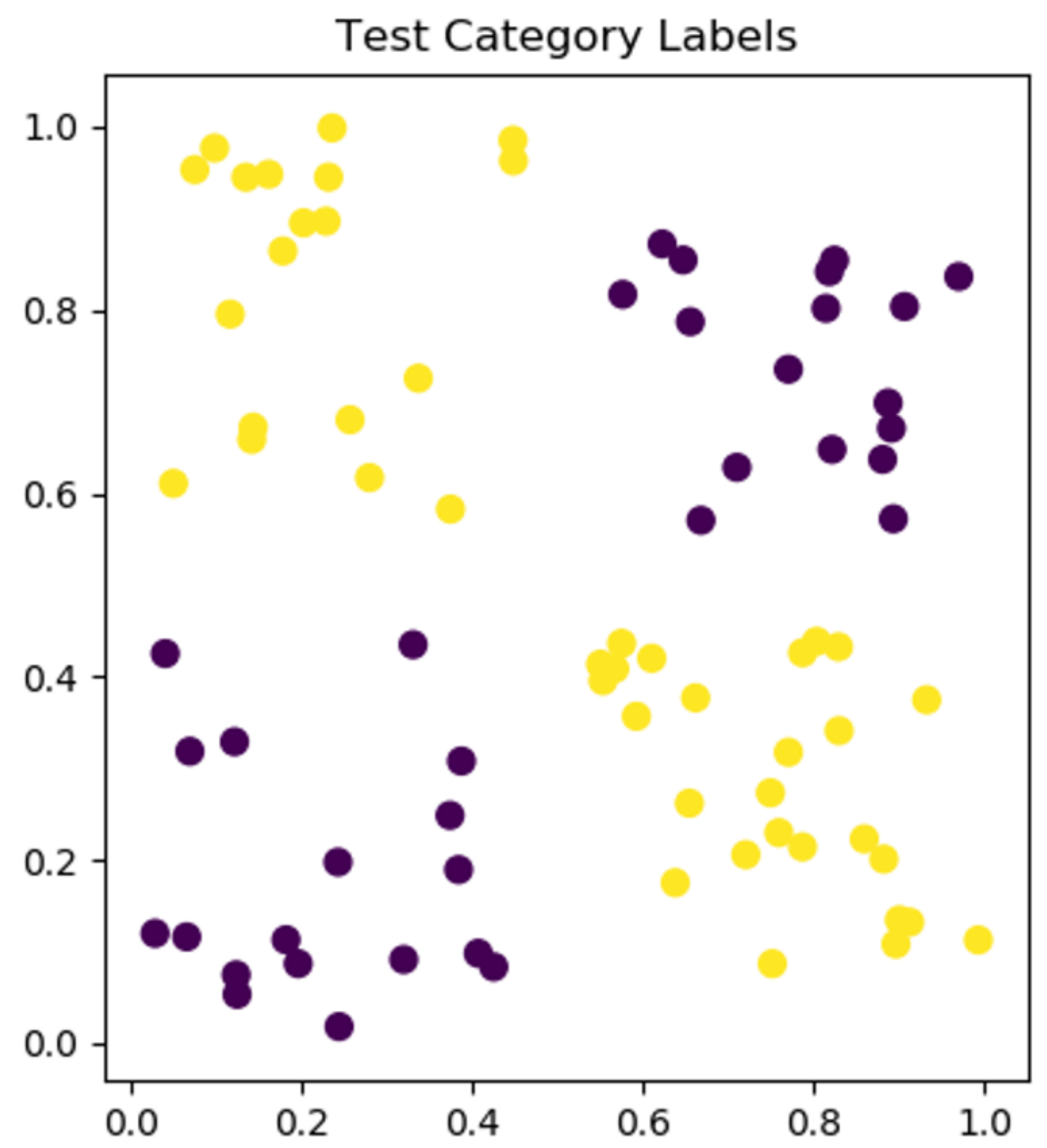


K-Means is performed on each category to create subcategories

Test Phase:



Test objects are classified into subcategories



Category label is deduced
Accuracy: 100%

ClusterCat Performance on datasets from the UCI Machine Learning Repository:

	XOR	Breast Cancer Wisconsin	Pima Diabetes	Contraceptive Method Choice	Iris	Cleveland Heart Disease
Average ClusterCat Accuracy	99.58% (SD = 0.07)	96.20% (SD = 1.63)	62.32% (SD = 2.97)	43.45% (SD = 2.28)	93.22% (SD = 4.33)	65.28% (SD =5.92)

Conclusion

- ClusterCat shows promise of improving supervised classification by utilizing unsupervised K-Means clustering to create subcategories to sort test objects into
- These results suggest that hierarchical categorization could be useful for supervised classification
- Although the results are promising, more research must be done to understand where and why ClusterCat gets certain classification decisions wrong

Limitations and Future Directions

- ClusterCat has a slightly slower runtime than scikit-learn's K-Means implementation
- Deeper examination into the datasets where ClusterCat succeeds may highlight cases where this algorithm could be most useful
- Examining if new point classification would be more accurate using an exemplar model rather than a prototypic one. K-Nearest Neighbor might be a useful algorithm to employ or take inspiration from to achieve this exemplar approach
- Humans often learn categorical information without supervision; leveraging unsupervised clustering in ML algorithms may improve learning

References

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.