

Lecture 14: 10 March, 2022

Madhavan Mukund

<https://www.cmi.ac.in/~madhavan>

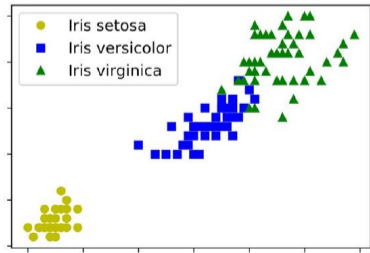
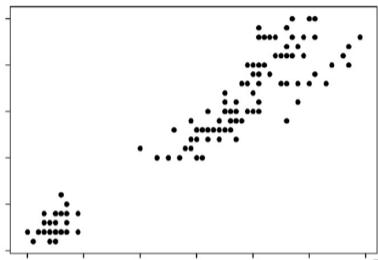
Data Mining and Machine Learning
January–May 2022

Unsupervised learning

- Supervised learning requires labelled data
- Vast majority of data is unlabelled
- What insights can you get into unlabelled data?

“If intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake ...”

- Yann LeCun
ACM Turing Award 2018



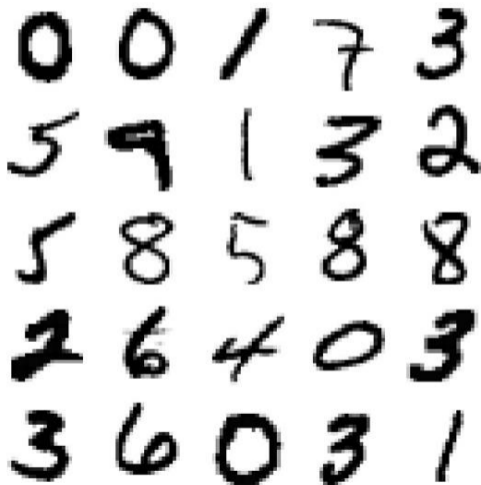
Applications

- Customer segmentation
 - Marketing campaigns
- Anomaly detection
 - Outliers
- Semi-supervised learning
 - Propagate limited labels
- Image segmentation
 - Object detection



Semi-supervised learning

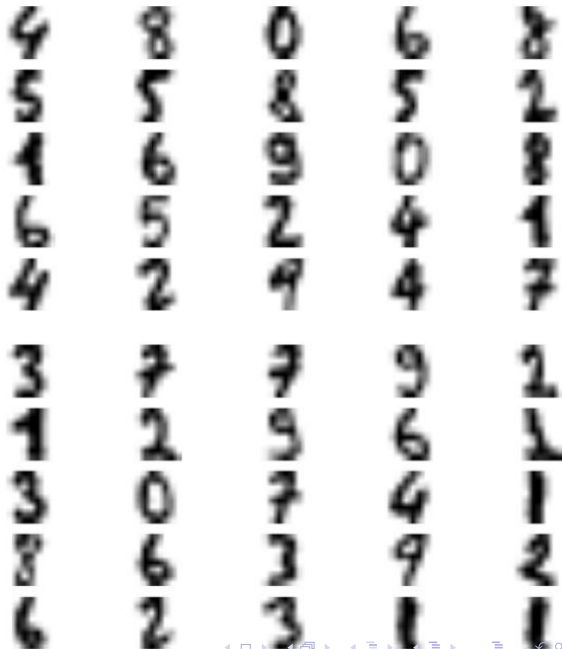
- Labelling training data is a bottleneck of supervised learning
- Handwritten digits 0,1,...,9
 - 1797 images
- Standard logistic regression model has 96.9% accuracy
- Suppose we take 50 random samples as training set
- Logistic regression gives 83.3%



Semi-supervised learning

- Instead of 50 random samples, 50 clusters using K means
- Use image nearest to each centroid as training set
 - 50 *representative images*
- Logistic regression accuracy jumps to 92.2%

97% → 83.3%



Semi-supervised learning



- Propagate representative image label to entire cluster
- Logistic regression improves to 93.3%
- Propagate representative image label to only 20% items closest to centroid
- Logistic regression improves to 94%
- Only 50 actual labels used, about 5 per class!

92.2% →

1797 → 50

360 ≈ 20%



Image segmentation

- An image is a matrix of pixels
- Each pixel has (R,G,B) values
- K means clustering on these values merges colours



Image segmentation

- An image is a matrix of pixels
- Each pixel has (R,G,B) values
- K means clustering on these values merges colours
- With 10 clusters, not much change

10 colors



Summary

- Unsupervised learning is useful as a preprocessing step
- Semi supervised learning
 - Identify a small subset of items to label manually
 - Propagate labels via cluster
- Image segmentation
 - Highlight objects by colour

0 0 1 7 3
5 9 1 3 2
5 8 5 8 8
2 6 4 0 3
3 6 0 3 1

