Lecture 14: 13 March, 2025

Madhavan Mukund

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

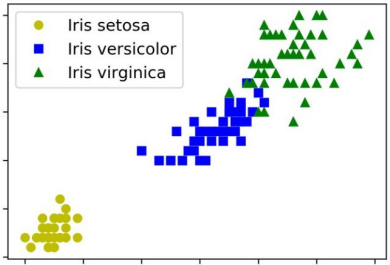
Data Mining and Machine Learning
January–April 2025

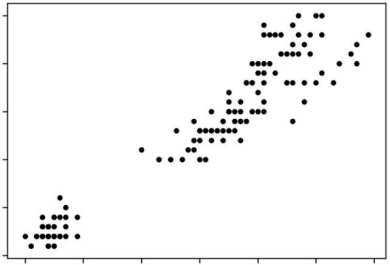
Unsupervised learning

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

Clustering for supervised learning

- Labelling training data is a bottleneck of supervised learning

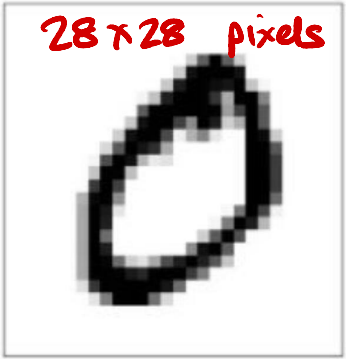
Clustering for supervised learning

- Labelling training data is a bottleneck of supervised learning
- Handwritten digits 0,1,. . . ,9
  - 1797 images
  - 8×8 pixels, grayscale
  - Each image is a 64-tuple $(x_1, x_2, \ldots, x_{64})$

MNIST

↓

digits 8x8



Clustering for supervised learning

- Labelling training data is a bottleneck of supervised learning
- Handwritten digits 0,1,. . . ,9
  - 1797 images
  - 8×8 pixels, grayscale
  - Each image is a 64-tuple $(x_1, x_2, \ldots, x_{64})$
- Standard logistic regression model has 96.9% accuracy

Clustering as preprocessing

- Use K Means to make 50 clusters

Clustering as preprocessing

- Use K Means to make 50 clusters
- Replace each input by its distance from the 50 centroids
  - Instead of $(x_1, x_2, \ldots, x_{64})$
  - $\ldots (d_1, d_2, \ldots, d_{50})$

Clustering as preprocessing

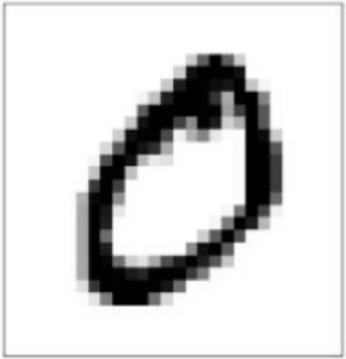
- Use K Means to make 50 clusters
- Replace each input by its distance from the 50 centroids
  - Instead of $(x_1, x_2, \ldots, x_{64})$
  - ... $(d_1, d_2, \ldots, d_{50})$
- Logistic regression on this representation jumps from 96.9% to 97.8% accuracy!

Clustering as preprocessing

- Use K Means to make 50 clusters
- Replace each input by its distance from the 50 centroids
  - Instead of $(x_1, x_2, \ldots, x_{64})$
  - ... $(d_1, d_2, \ldots, d_{50})$
- Logistic regression on this representation jumps from 96.9% to 97.8% accuracy!
- Varying the number of clusters changes the accuracy
  - 99 clusters is optimum, 98.2% accuracy

Semi-supervised learning

- 1797 images of handwritten digits 0,1,. . . ,9
- Standard logistic regression model has 96.9% accuracy



Semi-supervised learning

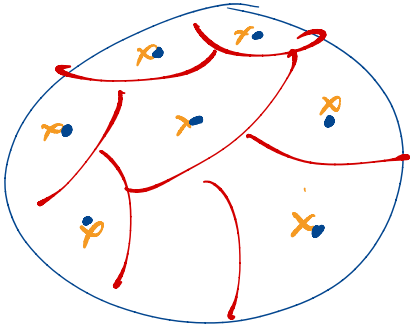
- 1797 images of handwritten digits 0,1,. . . ,9
- Standard logistic regression model has 96.9% accuracy
- What if we couldn't label the entire training set?

Semi-supervised learning

- 1797 images of handwritten digits 0,1,. . . ,9
- Standard logistic regression model has 96.9% accuracy
- What if we couldn't label the entire training set?
- Suppose we take 50 random samples as training set
- Logistic regression gives 83.3% accuracy

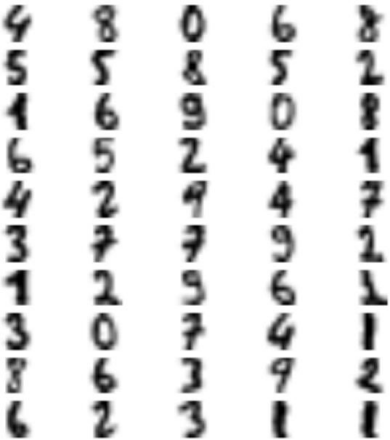
Semi-supervised learning

- Instead of 50 random samples, 50 clusters using K means
- Use image nearest to each centroid as training set



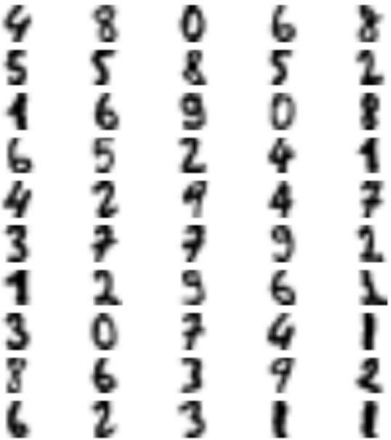
Semi-supervised learning

- Instead of 50 random samples, 50 clusters using K means
- Use image nearest to each centroid as training set
- 50 representative images
  - . . . but not randomly chosen 50



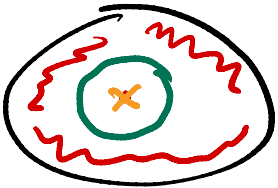
Semi-supervised learning

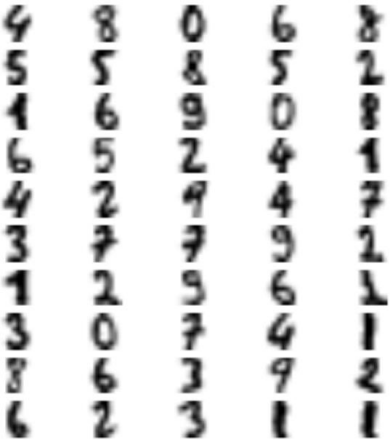
- Instead of 50 random samples, 50 clusters using K means
- Use image nearest to each centroid as training set
- 50 representative images
  - . . . but not randomly chosen 50
- Logistic regression accuracy jumps to 92.2%



Semi-supervised learning

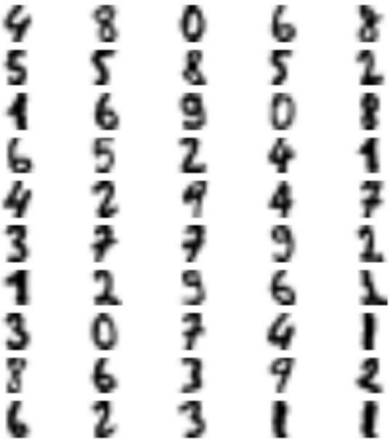
- Propagate representative image label to entire cluster
- Logistic regression improves to 93.3%





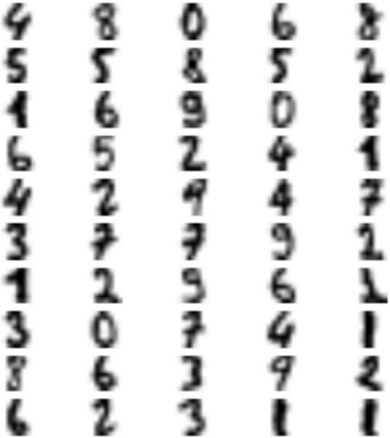
Semi-supervised learning

- Propagate representative image label to entire cluster
- Logistic regression improves to 93.3%
- Propagage representive image label to 20% items closest to centroid



Semi-supervised learning

- Propagate representative image label to entire cluster
- Logistic regression improves to 93.3%
- Propagage representive image label to 20% items closest to centroid
- Logistic regression improves to 94%



Semi-supervised learning

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

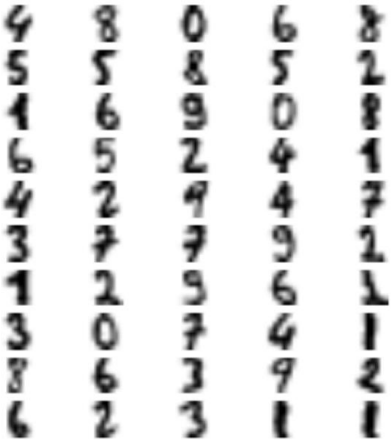


Image segmentation

- An image is a matrix of pixels
- Each pixel's colour is a triple (R,G,B) 0 - 255

Image segmentation

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

Image segmentation

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

Image segmentation

- An image is a matrix of pixels
- Each pixel's colour is a triple (R,G,B)
- K means clustering on these values merges colours
- With 10 clusters, not much change
- Same with 8

Image segmentation

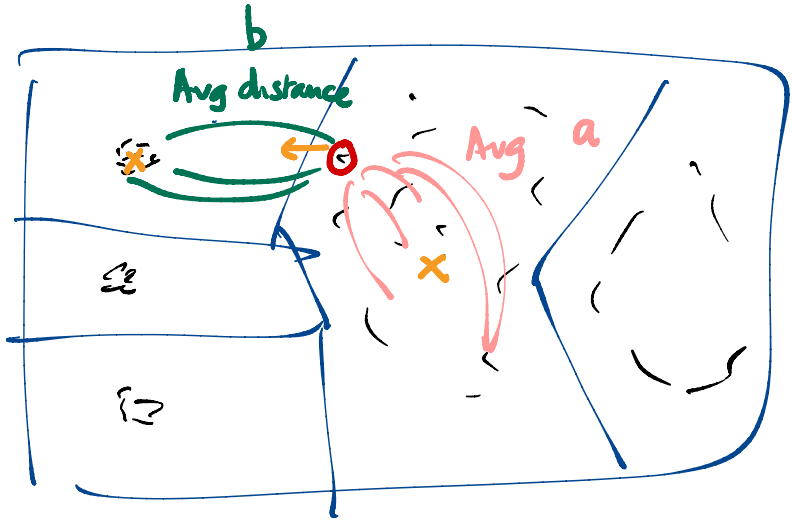
- An image is a matrix of pixels
- Each pixel's colour is a triple (R,G,B)
- K means clustering on these values merges colours
- With 10 clusters, not much change
- Same with 8
- At 6 colours, ladybug red goes

Image segmentation

- An image is a matrix of pixels
- Each pixel's colour is a triple (R,G,B)
- K means clustering on these values merges colours
- With 10 clusters, not much change
- Same with 8
- At 6 colours, ladybug red goes
- 4 colours

Image segmentation

- An image is a matrix of pixels
- Each pixel's colour is a triple (R,G,B)
- K means clustering on these values merges colours
- With 10 clusters, not much change
- Same with 8
- At 6 colours, ladybug red goes
- 4 colours
- Finally 2 colours, flower and rest



Good $b \gg a$

$\approx \frac{b}{b} = 1$

$= \frac{b-a}{\max(a,b)}$

$= \frac{-a}{a} = -1$

Bad case $a \gg b$