Lecture 15: 27 May, 2021

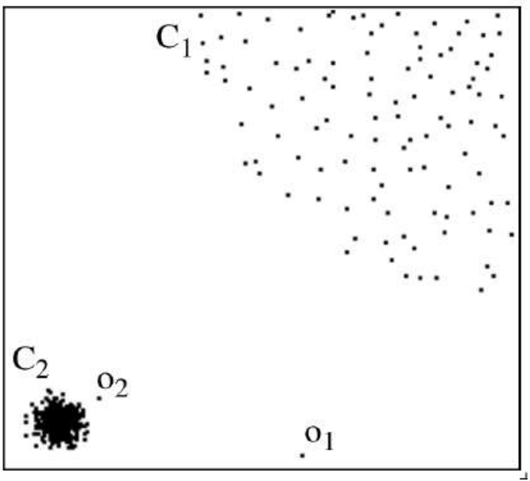
Madhavan Mukund

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

Data Mining and Machine Learning
April–July 2021

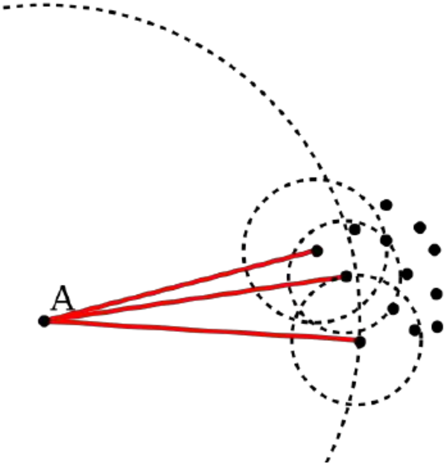
Outliers and density

- An outlier is less dense than its nearest neighbours
- But difference in density may be local
- A distance metric to eliminate o_2 could make all of C_1 outliers
- C_1 has 400 points, C_2 has 100 points
- Larger distance would make all of C_2 outliers with respect to C_1



Outliers and density

- For clustering, we defined a radius *Eps* and looked for *MinPts* neighbours within that ball
- Instead, fix *MinPts* and find smallest ball with that many neighbours
- Compare *radius(p)* with radius of its neighbours
- A is an outlier because its radius is much more than that of its neighbours

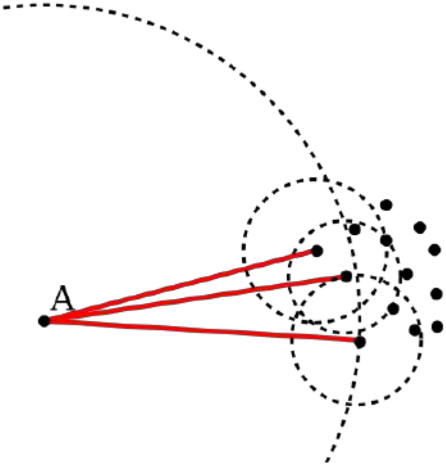


Outliers and density

- Local outlier factor *LOF(p)*

$$\frac{\text{Mean radius of } \mathit{MinPts\text{-}neighbours}(p)}{\mathit{radius}(p)}$$

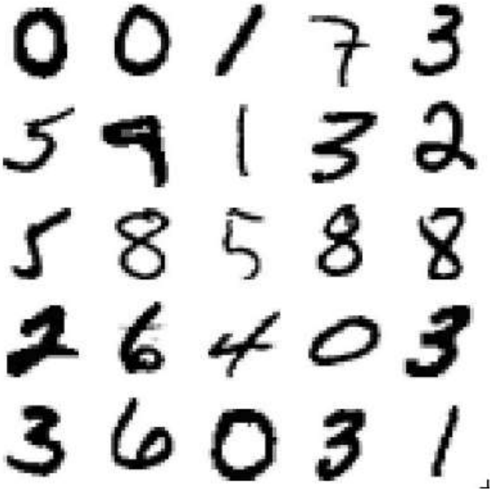
- The smaller this ratio, the more likely that *p* is an outlier
- Comparison is local to neighbourhood, so this can deal with different densities across range of data



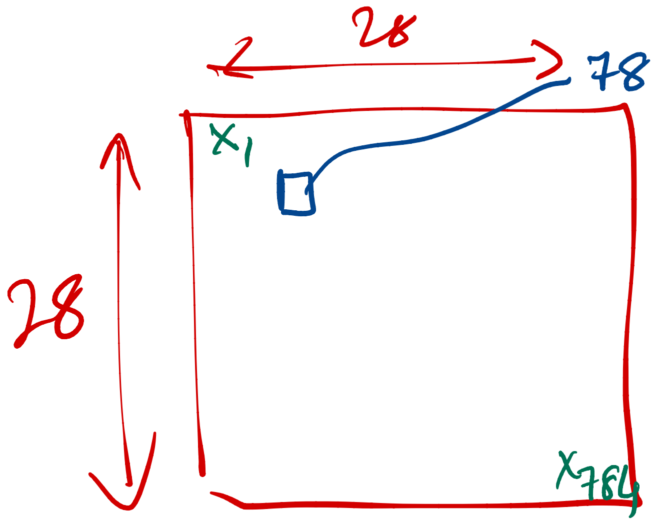
Semi-supervised learning

28×28 pixel gray scale

- 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%



Image



784 pixels

Each pixel is a grayscale value in $[0,1]$

Image is a vector

$(x_1, x_2, \dots, x_{784})$

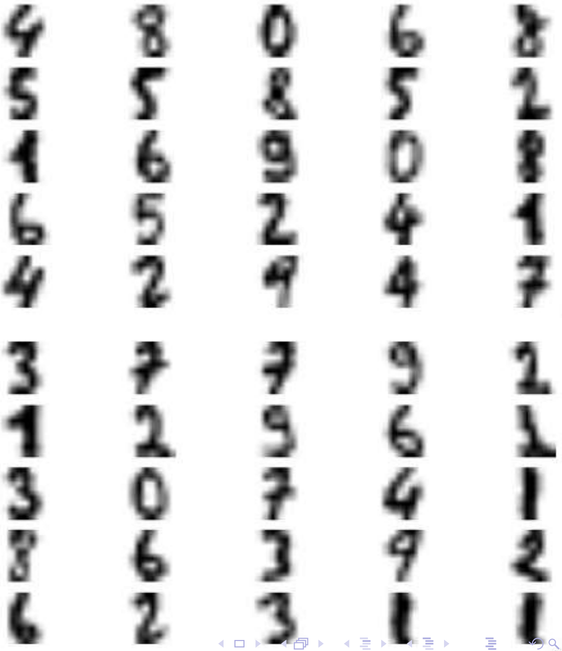
$\Big\{$ regression

$w_1x_1 + \dots + w_{784}x_{784}$

0 1 ... 9

Semi-supervised learning

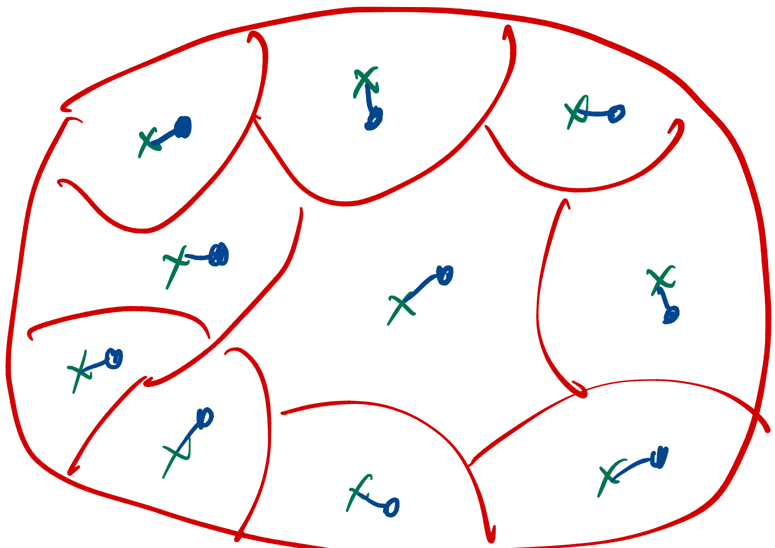
- Instead of 50 random samples, 50 clusters using K means
- Use image nearest to each centroid as training set
    - *50 representative images*
- Logistic regression accuracy jumps to 92.2%

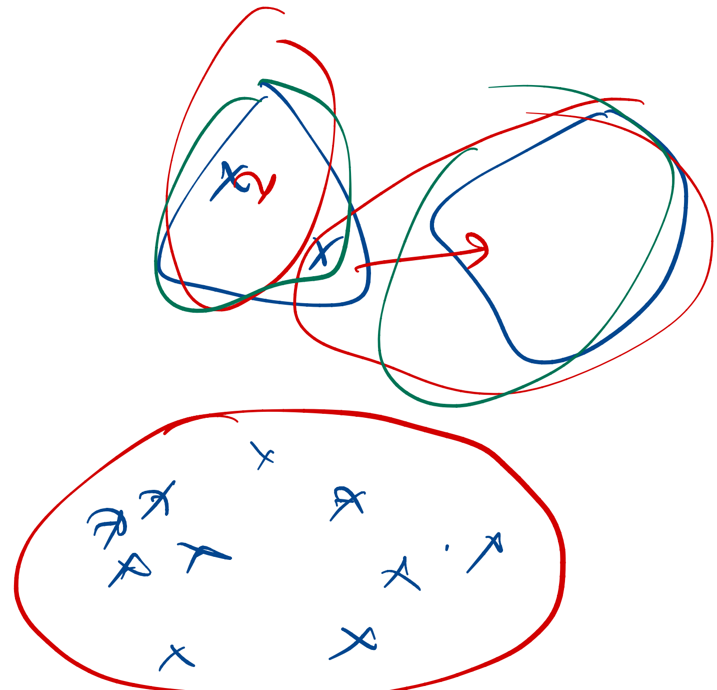


1800 images $(x_1, \ldots, x_{728})$

$(x_1', \ldots, x_{728}')$

$\downarrow$ K Means, $K = 50$





Semi-supervised learning

20% extrapolation

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

97% – Full training set

Random 50	K-50	K-50+
83%	92%	93

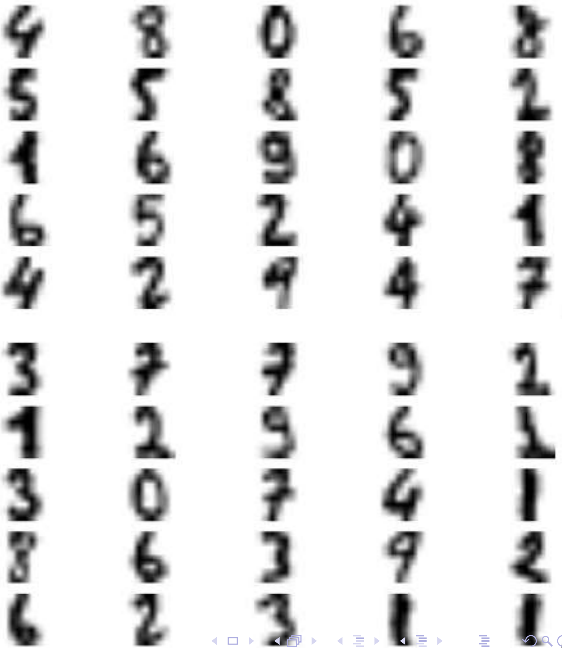


Image segmentation

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

2^8

$255 \times 255 \times 255$

$2^{24} = 2^4 \times 10^6$

Image segmentation

PNG → JPG

- 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

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
- Same with 8

8 colors

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
- Same with 8
- At 6 colours, ladybug red goes

6 colors

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
- Same with 8
- At 6 colours, ladybug red goes
- 4 colours

4 colors

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
- Same with 8
- At 6 colours, ladybug red goes
- 4 colours
- Finally 2 colours, flower and rest

2 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