Lecture 17: 7 June, 2021

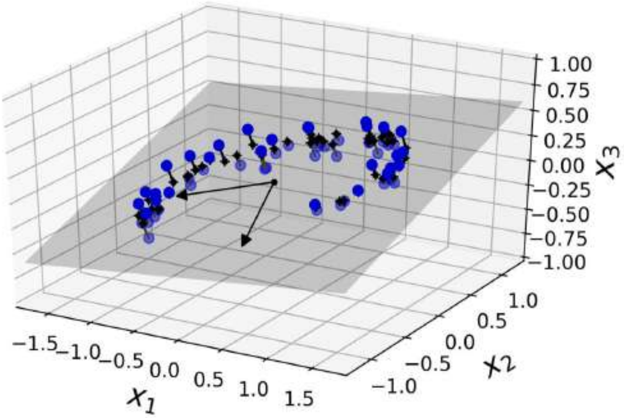
Madhavan Mukund

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

Data Mining and Machine Learning
April–July 2021

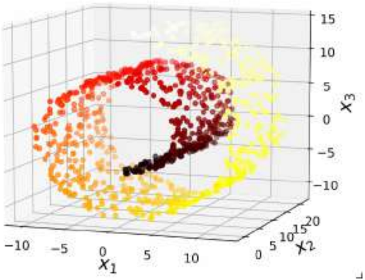
Dimensionality reduction

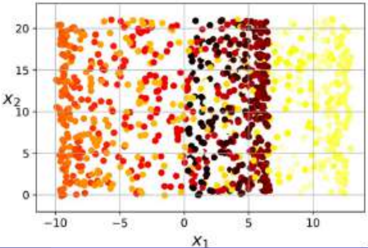
- Remove unimportant features by projecting to a smaller dimension
- Example: project blue points in 3D to black points in 2D plane
- Principal Component Anaylsis — transform d-dimensional input to k-dimensional input, preserving essential features
- Singular Value Decomposition (SVD)



Manifold learning

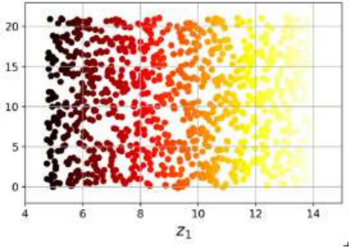
- Projection may not always help
- Swiss roll dataset
- Projection onto 2 dimesions is not useful



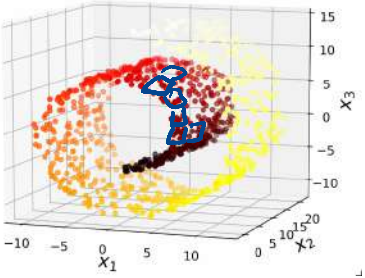


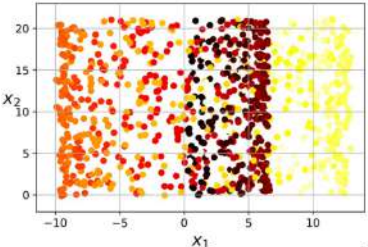
Manifold learning

- Projection may not always help
- Swiss roll dataset
- Projection onto 2 dimesions is not useful
- Better to unroll the image



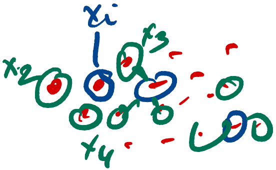
- Discover the manifold along which the data lies





Locally learning embeddings (LLE)

- Describe each point x_i as a linear combination of k nearest neighbours
  - Assume weight 0 for other neighbours



$$x_i = \sum_{j=1}^{n} w_{ij} x_j$$

only k nbrs have $w_{ij} \neq 0$

$k = 3$

$x_i = [x_{i1}, x_{i2} \ldots x_{im}]$

$$\begin{bmatrix} x_1 \\ \vdots \\ x_i \\ x_n \end{bmatrix} = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ w_{i1} & \cdots & w_{in} \\ w_{n1} & \cdots & w_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

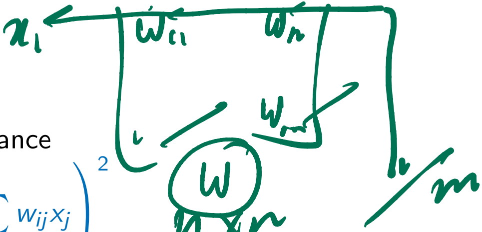
$$\begin{matrix} x_i & [w_{i1} \ldots w_{in}] \\ x_j & [w_{j1} \ldots w_{jn}] \end{matrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Locally ~~learning~~ Linear embeddings (LLE)

- Describe each point x_i as a linear combination of k nearest neighbours
    - Assume weight 0 for other neighbours

W=1

$$x_i = \sum_{j=1}^{n} w_{ij} x_j$$



- Choose weights to minimize the sum square distance

$$\widehat{W} = \arg\min_{W} \sum_{i=1}^{n} \Big(\underbrace{x_i}_{\text{original } x_i} - \underbrace{\sum_{j=1}^{n} w_{ij} x_j}_{\text{reconstruction of } x_i \text{ from nbrs via } W} \Big)^2$$

$X = W \cdot X$

Locally learning embeddings (LLE)

- Describe each point x_i as a linear combination of k nearest neighbours
  - Assume weight 0 for other neighbours

$$x_i = \sum_{j=1}^{n} w_{ij} x_j$$

- Choose weights to minimize the sum square distance

$$\widehat{W} = \arg\min_{W} \sum_{i=1}^{n} \left(x_i - \sum_{j=1}^{n} w_{ij} x_j \right)^2$$

- Captures "local" geometry
  - Already invariant with respect to rotation, scaling
  - Normalize weights to sum up to 1 — invariance under translation

$x + d$

$w_l d + w_j d$

$w_l \quad x_l + d$

$w_j \quad x_j + d$

Locally learning embeddings (LLE) . . .

- Original inputs are in m dimensions
- Map each x to a new vector z in $m' \ll m$ dimensions

Locally learning embeddings (LLE) . . .

- Original inputs are in m dimensions
- Map each x to a new vector z in $m' \ll m$ dimensions
- Choose new representation to preserve original weighted approximations

$$\hat{Z} = \arg\min_{Z} \sum_{i=1}^{\overset{n}{\cancel{m}}} \left(z_i - \sum_{j=1}^{\overset{n}{\cancel{m}}} w_{ij} z_j \right)^2$$

  - Solve using eigenvalues/eigenvectors of $\widehat{W}$

inherited from earlier step

$$\overset{m}{x_i} \longmapsto \overset{m'}{z_i}$$

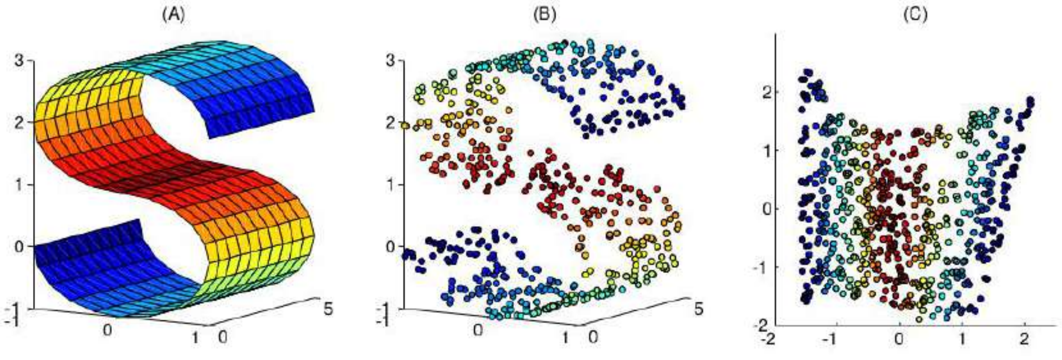
nbrs

x_{i1} $\quad$ z_{i1}

x_{i2} $\quad$ z_{i2}

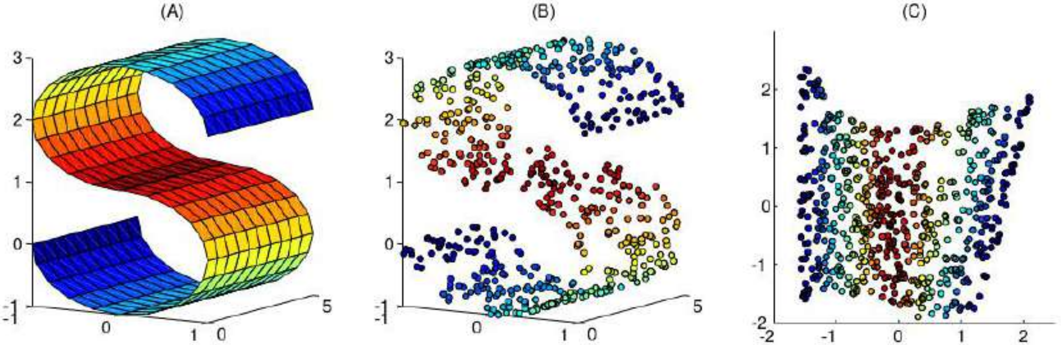
$\vdots$

Locally learning embeddings (LLE)



Original image | Sampled points | LLE reconstruction

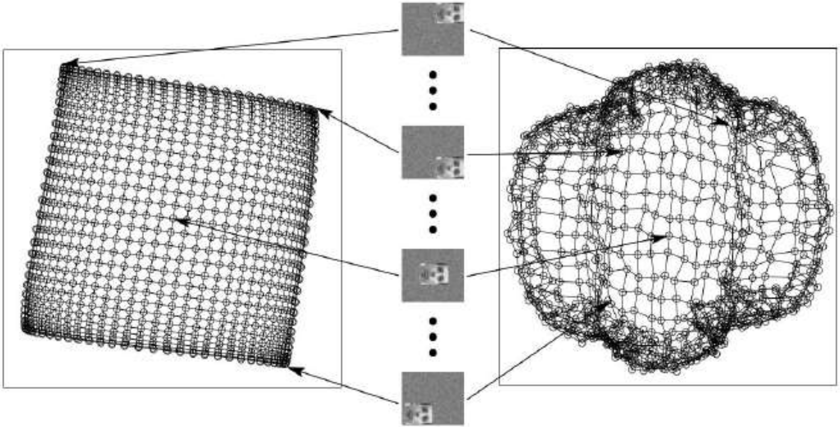
Locally learning embeddings (LLE)



Original image | Sampled points | LLE reconstruction

- Need enough samples to discover the "curves"

Locally learning embeddings (LLE)



LLE reconstruction preserves neighbourhood structure

PCA distorts geometry

Mixture models

- Probabilistic process — parameters Θ
  - Tossing a coin with $\Theta = \{Pr(H)\} = \{p\}$

Mixture models

- Probabilistic process — parameters Θ
  - Tossing a coin with $\Theta = \{Pr(H)\} = \{p\}$
- Perform an experiment
  - Toss the coin N times, $H \ T \ H \ H \ \cdots \ T$

Mixture models

- Probabilistic process — parameters Θ
  - Tossing a coin with $\Theta = \{Pr(H)\} = \{p\}$
- Perform an experiment
  - Toss the coin N times, $H\ T\ H\ H\ \cdots\ T$
- Estimate parameters from observations
  - From h heads, estimate $p = h/N$
  - Maximum Likelihood Estimator (MLE)

Mixture models

- Probabilistic process — parameters Θ
  - Tossing a coin with $\Theta = \{Pr(H)\} = \{p\}$
- Perform an experiment
  - Toss the coin N times, $H\ T\ H\ H\ \cdots\ T$
- Estimate parameters from observations
  - From h heads, estimate $p = h/N$
  - Maximum Likelihood Estimator (MLE)
- What if we have a mixture of two random processes

Mixture models

- Probabilistic process — parameters Θ
  - Tossing a coin with $\Theta = \{Pr(H)\} = \{p\}$
- Perform an experiment
  - Toss the coin N times, $H\ T\ H\ H\ \cdots\ T$
- Estimate parameters from observations
  - From h heads, estimate $p = h/N$
  - Maximum Likelihood Estimator (MLE)
- What if we have a mixture of two random processes
  - Two coins, c_1 and c_2, with $Pr(H) = p_1$ and p_2, respectively

Mixture models

- Probabilistic process — parameters Θ
  - Tossing a coin with $\Theta = \{Pr(H)\} = \{p\}$
- Perform an experiment
  - Toss the coin N times, $H\ T\ H\ H\ \cdots\ T$
- Estimate parameters from observations
  - From h heads, estimate $p = h/N$
  - Maximum Likelihood Estimator (MLE)
- What if we have a mixture of two random processes
  - Two coins, c_1 and c_2, with $Pr(H) = p_1$ and p_2, respectively
  - Repeat N times: choose c_i with probability $1/2$ and toss it

Mixture models

- Probabilistic process — parameters Θ
  - Tossing a coin with $\Theta = \{Pr(H)\} = \{p\}$
- Perform an experiment
  - Toss the coin N times, $H\ T\ H\ H\ \cdots\ T$

HHTTHTHHHTTHTH

- Estimate parameters from observations
  - From h heads, estimate $p = h/N$
  - Maximum Likelihood Estimator (MLE)
- What if we have a mixture of two random processes
  - Two coins, c_1 and c_2, with $Pr(H) = p_1$ and p_2, respectively
  - Repeat N times: choose c_i with probability $1/2$ and toss it
  - Outcome: N_1 tosses of c_1 interleaved with N_2 tosses of c_2, $N_1 + N_2 = N$

Mixture models

- Probabilistic process — parameters Θ
  - Tossing a coin with $\Theta = \{Pr(H)\} = \{p\}$
- Perform an experiment
  - Toss the coin N times, $H\ T\ H\ H\ \cdots\ T$
- Estimate parameters from observations
  - From h heads, estimate $p = h/N$
  - Maximum Likelihood Estimator (MLE)
- What if we have a mixture of two random processes
  - Two coins, c_1 and c_2, with $Pr(H) = p_1$ and p_2, respectively
  - Repeat N times: choose c_i with probability $1/2$ and toss it
  - Outcome: N_1 tosses of c_1 interleaved with N_2 tosses of c_2, $N_1 + N_2 = N$
  - Can we estimate p_1 and p_2?

Mixture models . . .

- Two coins, c_1 and c_2, with $Pr(H) = p_1$ and p_2, respectively
- Sequence of N interleaved coin tosses $H\ T\ H\ H\ \cdots\ H\ H\ T$

Mixture models . . .

- Two coins, c_1 and c_2, with $Pr(H) = p_1$ and p_2, respectively
- Sequence of N interleaved coin tosses $H\ T\ H\ H\ \cdots\ H\ H\ T$
- If the sequence is labelled, we can estimate p_1, p_2 separately
  - $H\ T\ T\ H\ H\ T\ H\ T\ H\ H\ T\ H\ T\ H\ T\ H\ H\ T\ H\ T$
  - $p_1 = 8/12 = 2/3$, $p_2 = 3/8$

Mixture models . . .

- Two coins, c_1 and c_2, with $Pr(H) = p_1$ and p_2, respectively
- Sequence of N interleaved coin tosses $H\ T\ H\ H\ \cdots\ H\ H\ T$
- If the sequence is labelled, we can estimate p_1, p_2 separately
  - $H\ T\ T\ H\ H\ T\ H\ T\ H\ H\ T\ H\ T\ H\ T\ H\ H\ T\ H\ T$
  - $p_1 = 8/12 = 2/3$, $p_2 = 3/8$
- What the observation is unlabelled?
  - $H\ T\ T\ H\ H\ T\ H\ T\ H\ H\ T\ H\ T\ H\ T\ H\ H\ T\ H\ T$

Mixture models ...

- Two coins, c_1 and c_2, with $Pr(H) = p_1$ and p_2, respectively
- Sequence of N interleaved coin tosses $H\ T\ H\ H\ \cdots\ H\ H\ T$
- If the sequence is labelled, we can estimate p_1, p_2 separately
  - $H\ T\ T\ H\ H\ T\ H\ T\ H\ H\ T\ H\ T\ H\ T\ H\ H\ T\ H\ T$
  - $p_1 = 8/12 = 2/3$, $p_2 = 3/8$
- What the observation is unlabelled?
  - $H\ T\ T\ H\ H\ T\ H\ T\ H\ H\ T\ H\ T\ H\ T\ H\ H\ T\ H\ T$
- Iterative algorithm to estimate the parameters
  - Make an initial guess for the parameters
  - Compute a (fractional) labelling of the outcomes
  - Re-estimate the parameters

$d_1 \quad t_1$

$d_2 \quad t_2$

$\vdots \quad \vdots$

$d_N \quad t_N$

$\left.\begin{array}{l} P(t) \\ P(w|t) \end{array}\right\} P(d|t)$