University of Nottingham
UK | CHINA | MALAYSIA

COMP3009: Machine Learning (MLE)

- Feature Selection and Dimensionality Reduction

Dr Xin Chen
School of Computer Science
University of Nottingham

Aim of FS and DR

- Reduce the impact caused by Curse of Dimensionality
- Remove redundant features to improve performance
- Increase computational efficiency
- Reduce cost in new data acquisition
- FS vs DR
    - FS retain a subset of the original features
    - DR generate a new set of features that is compact but does retain the original meaning of features.

Things to consider when using FS and DR

- The target dimension
- Interpretability (Yes: FS; No: Dr or FS)
- Feature correlations/dependency
- Feature reliability and repeatability
- Methods (different methods likely to result in different features)

Popular FS Methods

- Wrapper methods
  - Search for optimal feature subset that maximise the decision-making performance
  - Methods: recursive feature elimination; sequential feature selection.
- Embedded methods
  - Integrate the FS process to the model learning process.
  - Methods: ridge (ElasticNet); lasso; random forest (feature ranking).
- Filter-based methods
  - Selection is based on feature relationships and statistics rather than performance.
  - Methods: univariate (ANOVA); Chi Square; correlation/variance

Forward Feature Selection (Wrapper Method)

Let $F = \{F_1, \ldots, F_n\}$ be the pool of potential features and let $M(X)$ be the evaluation metric for feature set X.

$X \leftarrow \emptyset$
while $X \neq F$
    $B \leftarrow 0$
    $Y \leftarrow \emptyset$
    for each $X_i \in F \setminus X$
        if $M(X \cup \{X_i\}) > B$ **then**
            $B \leftarrow M(X \cup \{X_i\})$
            $Y \leftarrow X \cup \{X_i\}$
    if $M(X) > B$ **then**
        return X
    else
        $X \leftarrow Y$
return X

Features that belongs to F not X

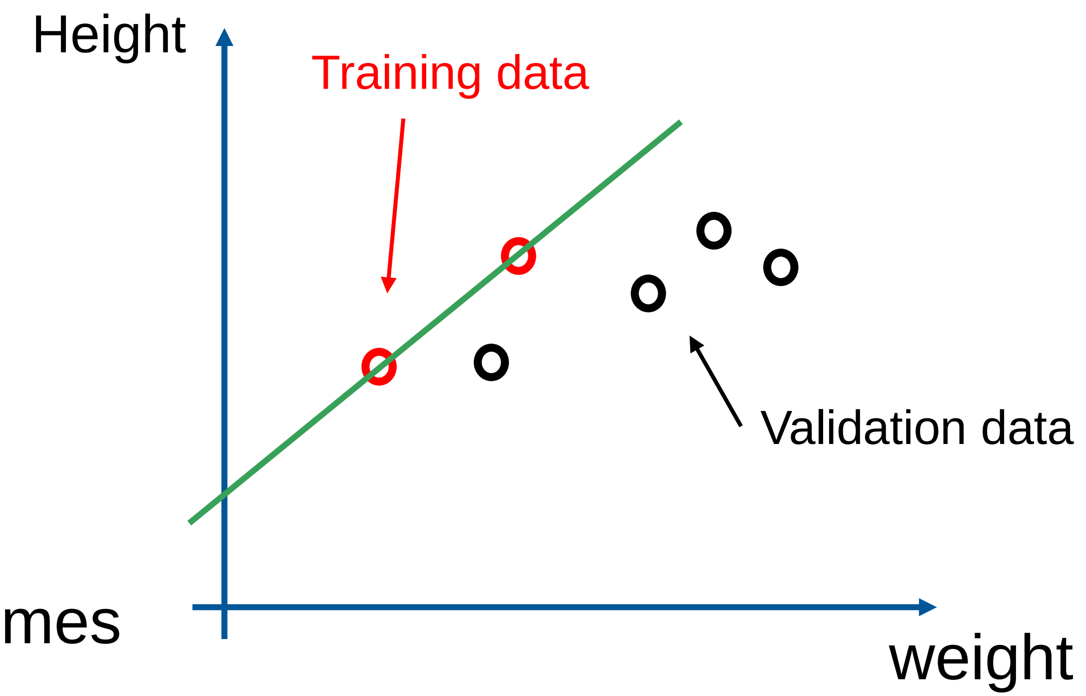
- **X**: the final selected feature set
- **B**: is stored best evaluation metric value
- **Y:** the selected feature set at each iteration
- **M:** The evaluation metric, e.g. Entropy; Classification rate; Regression error
- **Recursive feature elimination** method is similar: starts with the full set and eliminate one at a time.

LASSO (Embedded Method)

- LASSO (least absolute shrinkage and selection operator)
- Add a L1 regularisation term to reduce the number of effective features
- The loss function is not differentiable. Sub-gradient methods or least-angle regression can be used to optimise the loss

$$\min_{x}(\parallel Ax - b \parallel^2 + \lambda|x|)$$

Least squared term   L1 regularisation term



For multiple features:
e.g. $y=x_0+x_1*a1+x_2*a2+x_3*a3$

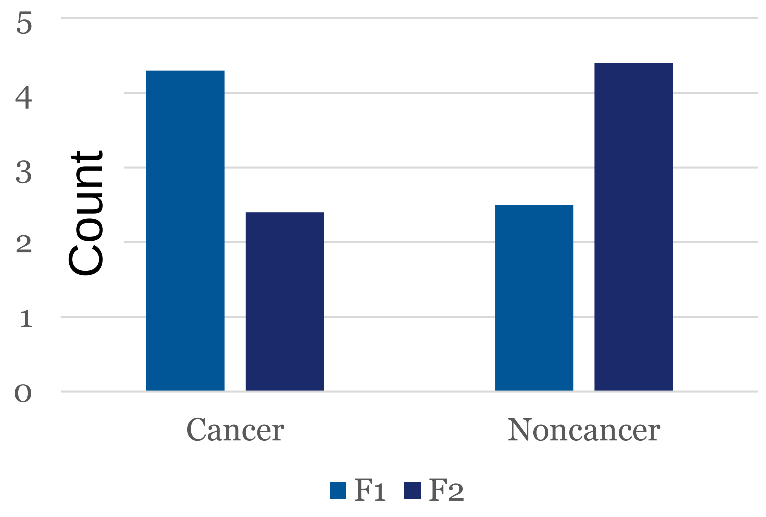
LASSO a higher λ will make some of x becomes 0, hence reduce the dimensionalities.

Chi Square vs T-test vs ANOVA (Filter Method)

- Univariant feature selection (assuming features are independent to each other)
- A chi-square tests the independence of predictor and outcome event, suitable for categorical features in categorical outcome.
- T-test compares the statistical difference of two groups (binary class) and used for continues features
- ANOVA uses variance to test the relationship between categorical predictors and continuous outcome response (e.g. gender, age to predict exam mark).
- Correlation test work for predictors and outcome are both continuous.
- Assume a null-hypothesis. Use p value to reject the null-hypothesis (e.g. $p<0.01$). P value indicates the probability of the null-hypothesis is true.

Chi Square

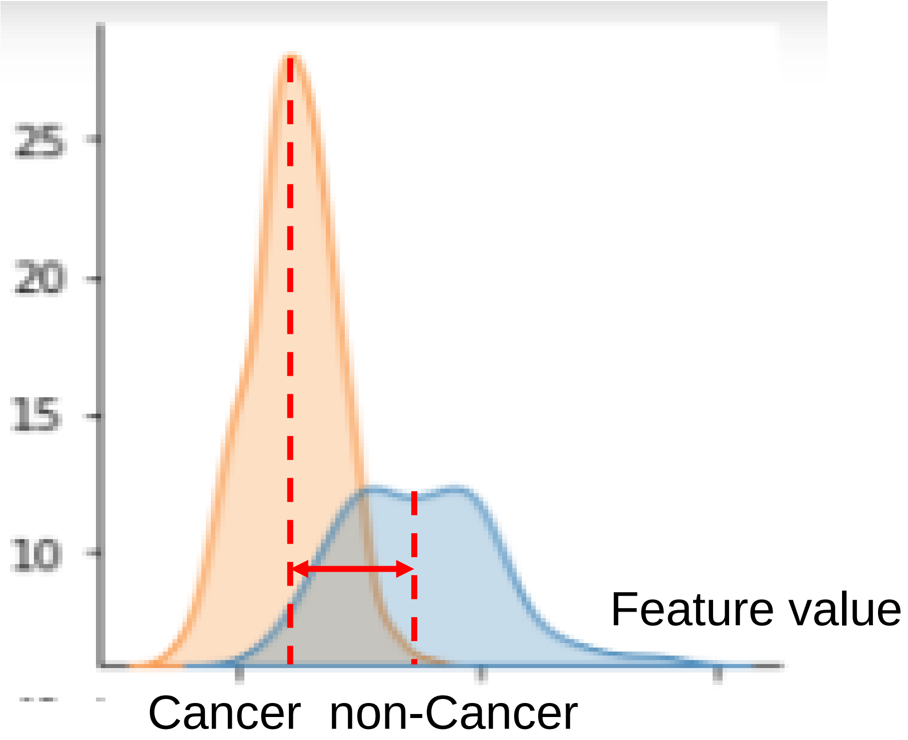
How categorical predictor associated with the categorical outcomes



null-hypothesis: the two categorical data are independent

T-test

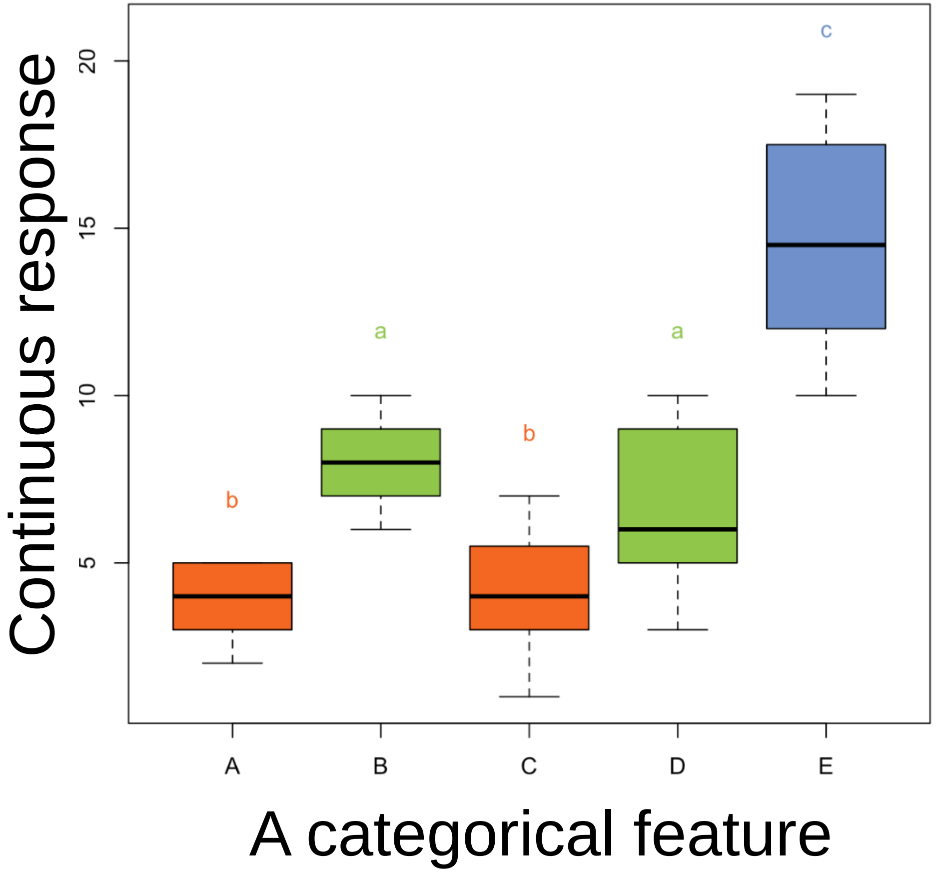
How the mean of two groups are different from each other



null-hypothesis: the mean of the two groups are the same

ANOVA

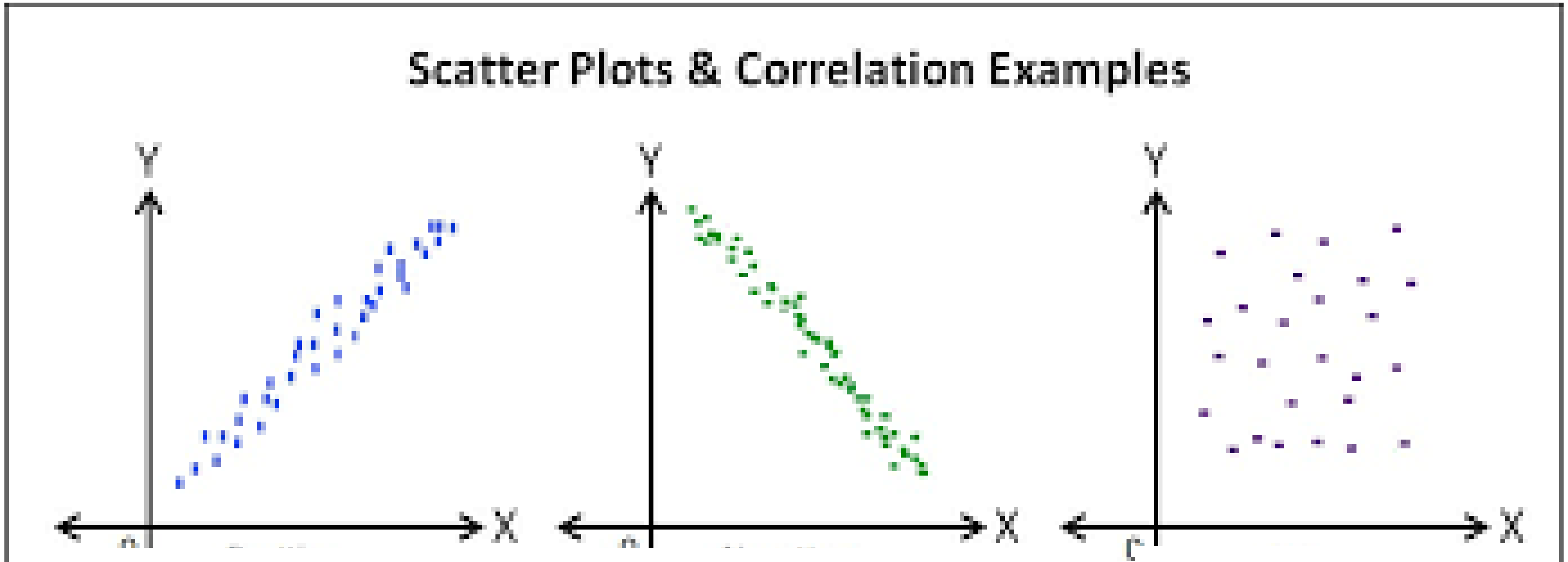
ANOVA checks variance between groups of categorical feature with respect to continuous response. Variance between the groups and variance within the groups.



null-hypothesis: the variance across categories are equal.

Correlation

Correlation of continuous predictor and continuous outcome (e.g. Pearson correlation coefficient)



Positive Correlation (close to 1)

Negative Correlation (close to -1)

No Correlation (close to 0)

null-hypothesis: the two group of data are not correlated.

Popular DR Methods

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Manifold learning (non-linear)

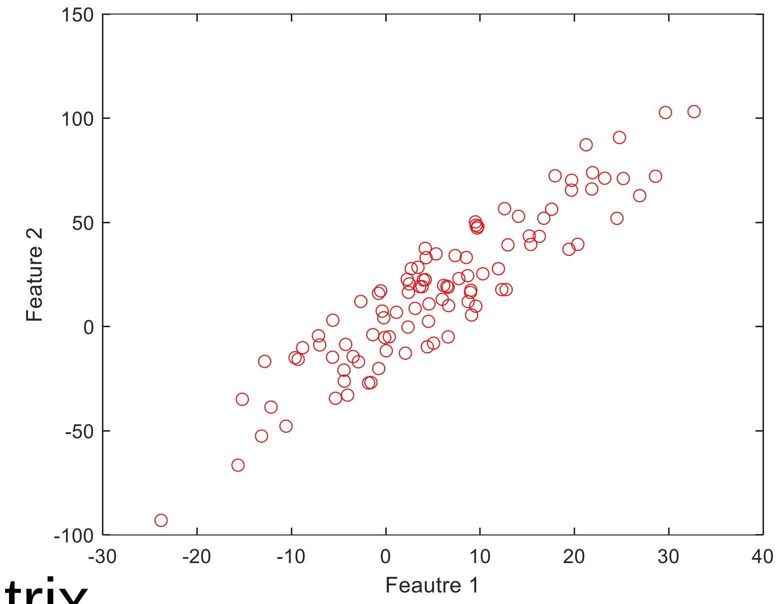
Principal Component Analysis

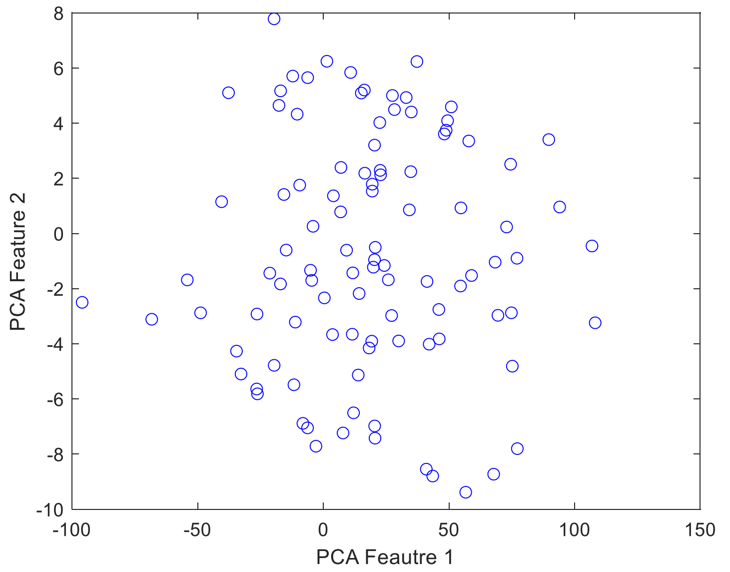
- Two methods resulted the same PCA calculation.
  - Maximum variance
  - Minimise average projection error

- PCA requires calculation of:
  - Mean of observed variables
  - Covariance of observed variables
  - Eigenvalue/eigenvector computation of covariance matrix

- How to calculate Eigen vectors and Eigen values:
  https://www.youtube.com/watch?v=TQvxWaQnrqI

- Demo





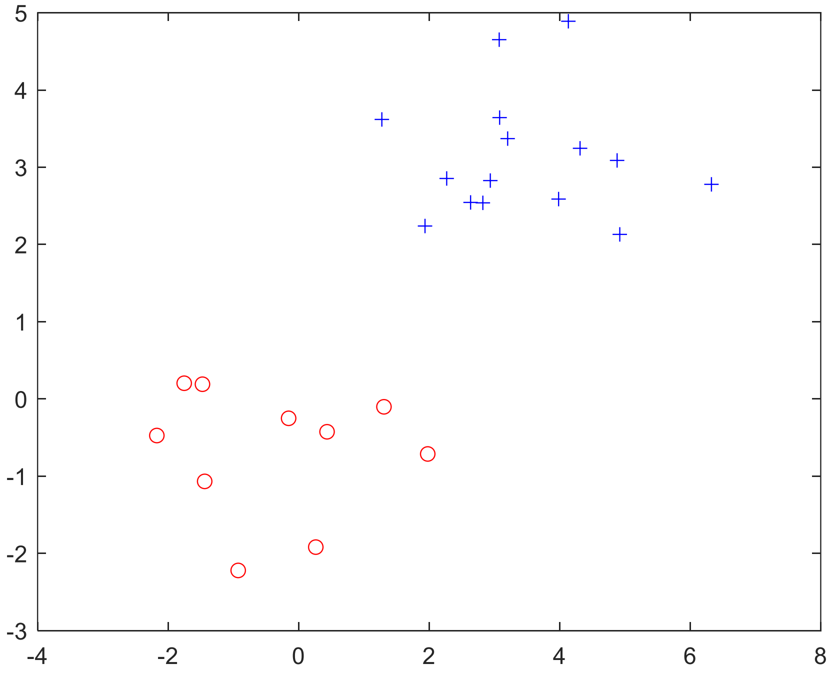
Linear Discriminant Analysis

- LDA is a predictive modelling algorithm for multi-class classification.
- Dimensionality reduction by providing a projection of a training dataset that best separates the examples by their assigned class.
- PCA :unsupervised; LDA: supervised
- How multi-class works?

Class mean

$$J(\boldsymbol{w}) = \frac{m_2 - m_1}{s_1^2 + s_2^2}$$

Class variance



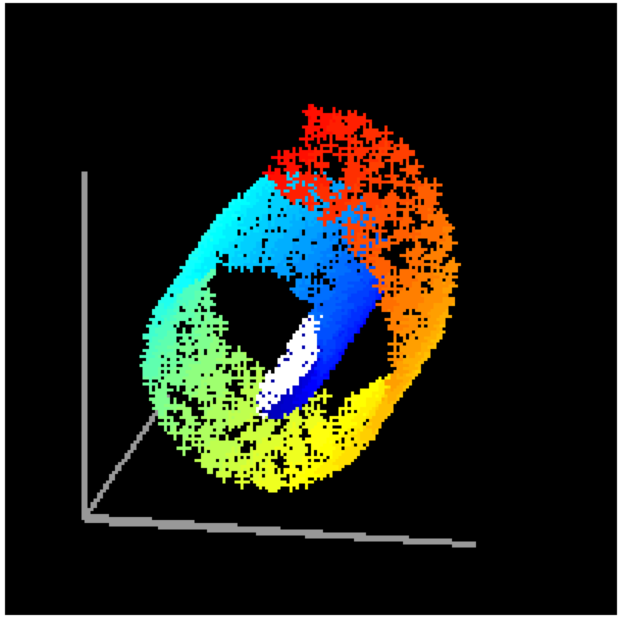
Manifold Learning

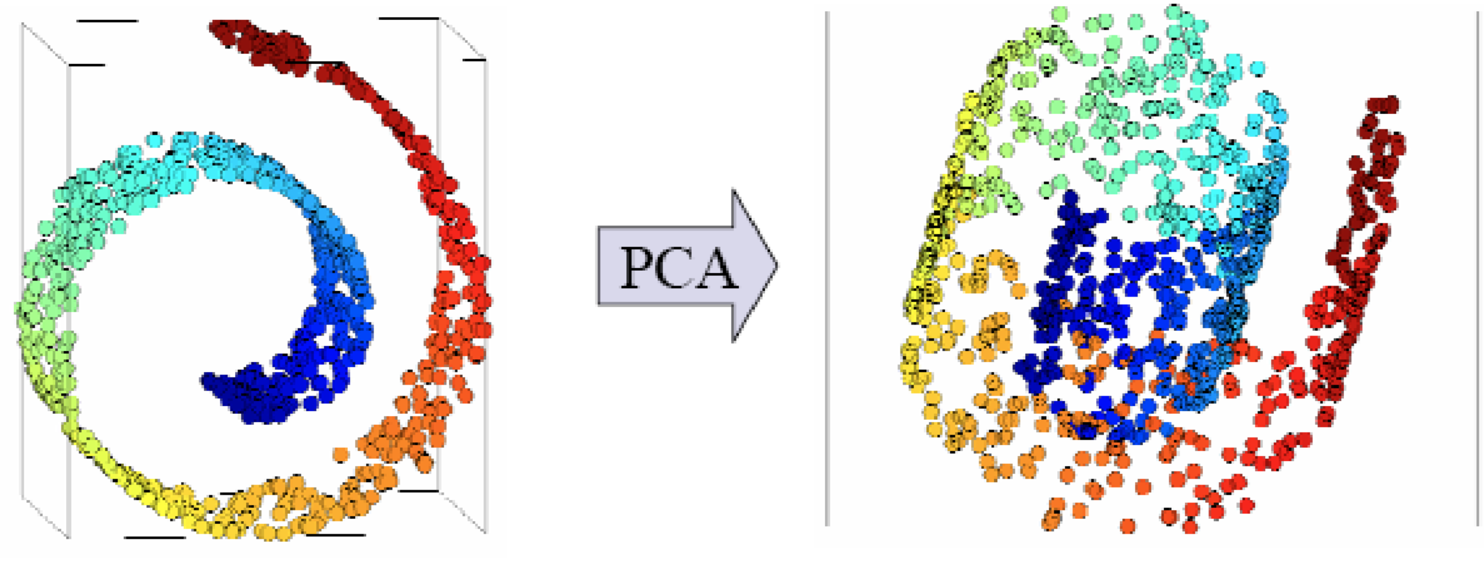
- The intuition: high-dimensional datasets often vary due to only a small number of parameters.
- E.g. dataset of 25, 64x64 images:

- Dataset actually varies due to only parameters: 2 angles + 1 illumination
- The images span a 3-dimensional manifold of R^{4096}

Manifold Learning

- Manifold learning aims to learn the latent representation of the original data in lower dimensions.
- PCA is a linear manifold learning method, which doesn't work in some cases.
- Swiss roll: what is the minimum dimension to present the data?



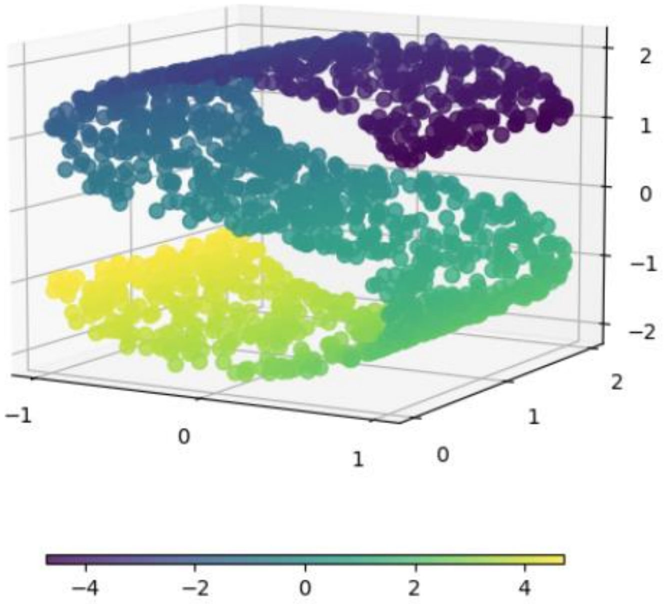


3D

2D

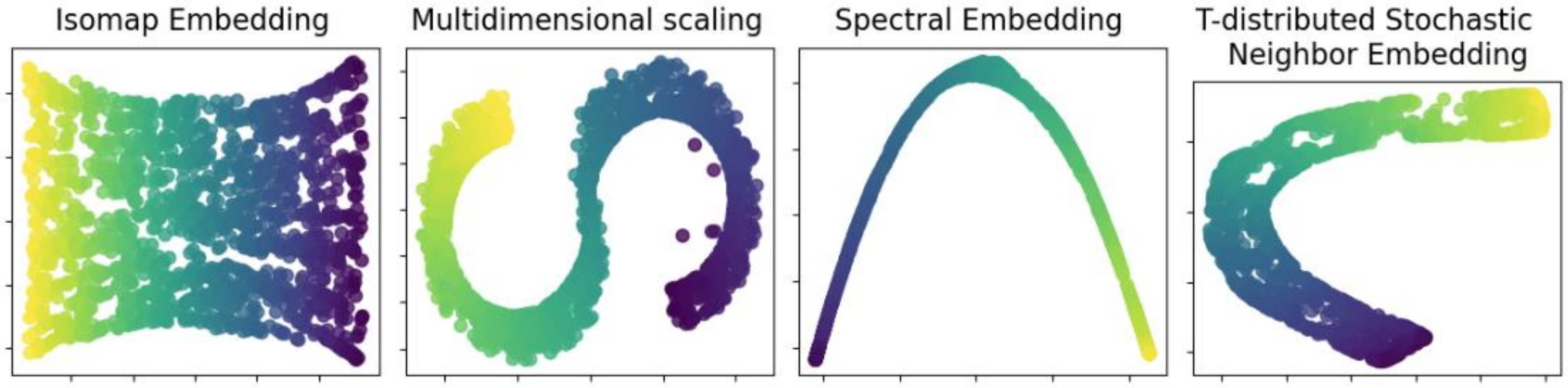
Nonlinear Manifold Learning Methods

Original data in 3D



Similar ideas to preserve local neighbouring relationships

Learned Manifold in 2D



Key Messages

- Understand the difference of FS and DR
- Know when to FS and DR
- Understand the working principle of key FS methods
- Understand the working principle of PCA & LDA
- Understand the concept of manifold learning
- Know some of the non-linear manifold learning methods