Computational Data Analysis

Machine Learning

Yao Xie, Ph.D.

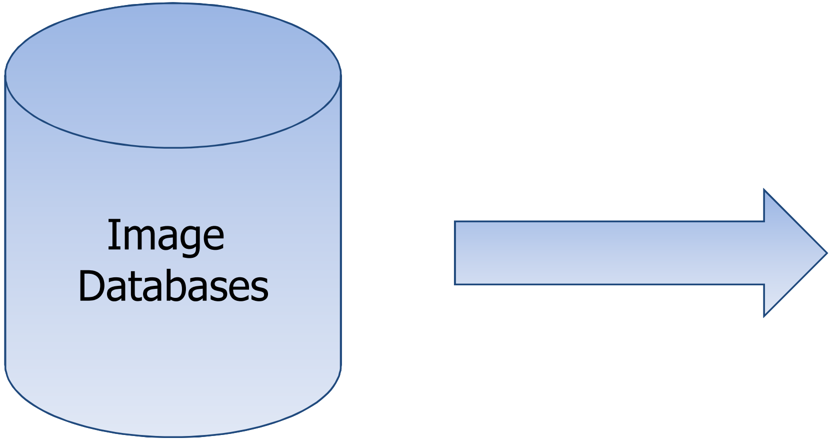
Associate Professor

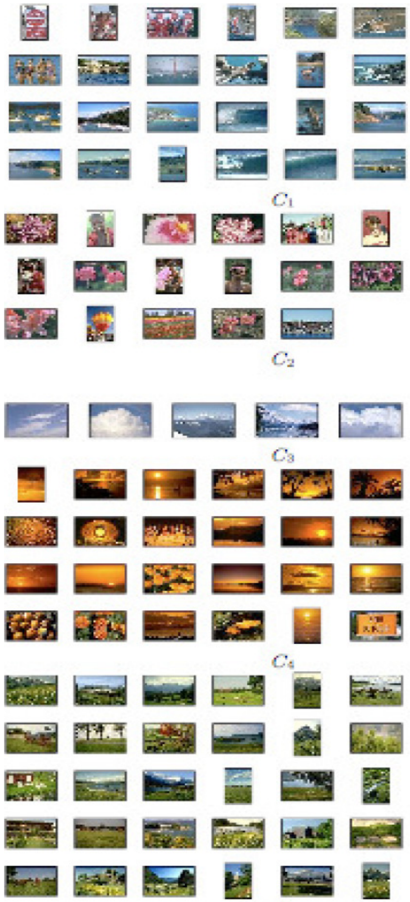
Harold R. and Mary Anne Nash Early Career Professor

H. Milton Stewart School of Industrial and Systems Engineering

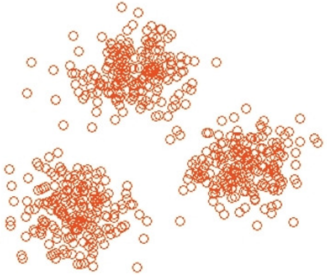
Dimensionality Reduction

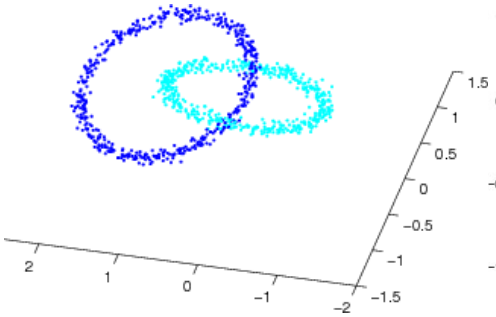
Principal Component Analysis

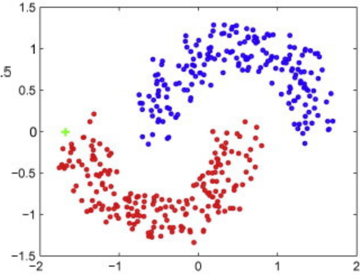




What are the relations between data points?

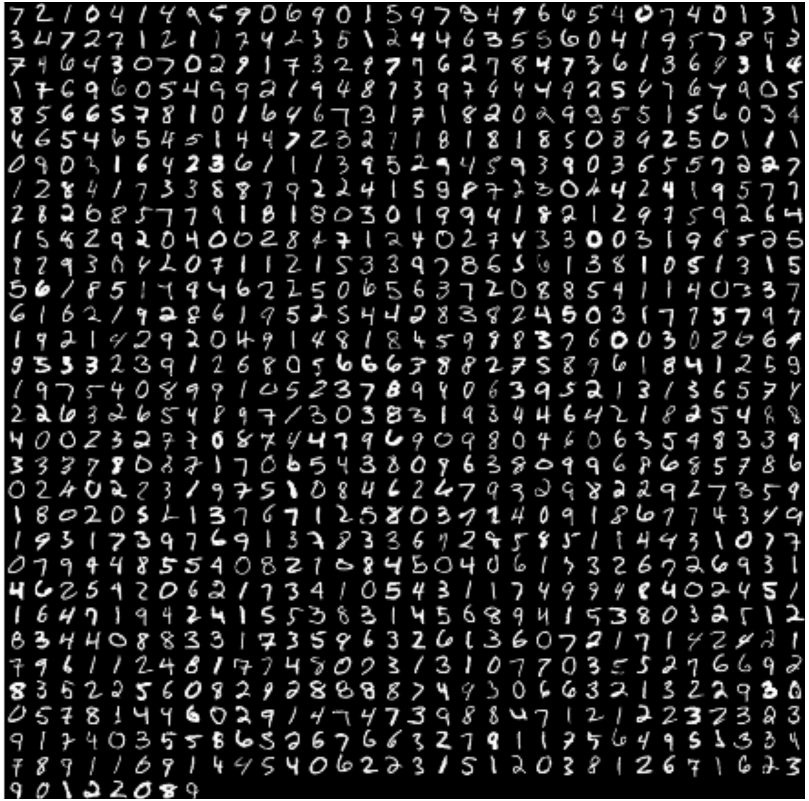


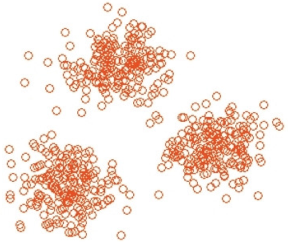


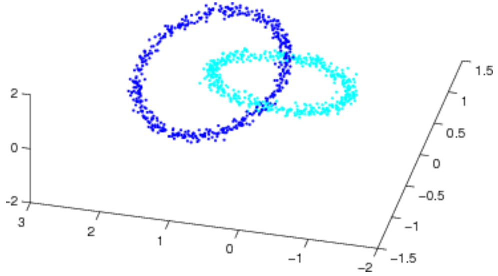


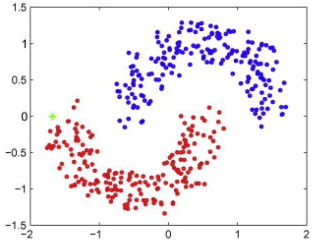
Handwritten digits

What are the relations between data points?





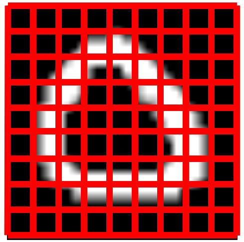




http://www.datawrangling.org/python-montage-code-for-displaying-arrays/

What is dimensionality reduction?

- The process of reducing the number of random variables under consideration
  - One can combine, transform or select variables
  - One can use linear or nonlinear operations



$$\begin{pmatrix}0\\0\\0\\1\\1\\1\\0\\0\\0\end{pmatrix}$$

vector in R^d

Original data point

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \xrightarrow{f(x): R^d \mapsto R^k} \quad z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{pmatrix}$$

Reduced representation

$$k \ll d$$

Why dimensionality reduction and how to think

- The dimension-reduced data can be used for
  - Visualizing, exploring and understanding the data
  - Aggregating weak signals in the data
  - Cleaning the data
  - Speeding up subsequent learning task
  - Building simpler model later

- Key questions of a dimensionality reduction algorithm
  - What is the criterion for carrying out the reduction process?
  - What are the algorithm steps?

Principal component analysis

- Given m data points, $\{x^1, x^2, \dots x^m\} \in R^d$, with mean

- Step 1: Estimate the mean and covariance matrix from data

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^i \quad and \quad C = \frac{1}{m}\sum_{i=1}^{m}(x^i - \mu)(x^i - \mu)^{\top}$$

"weights" to combine features: **Principal directions**

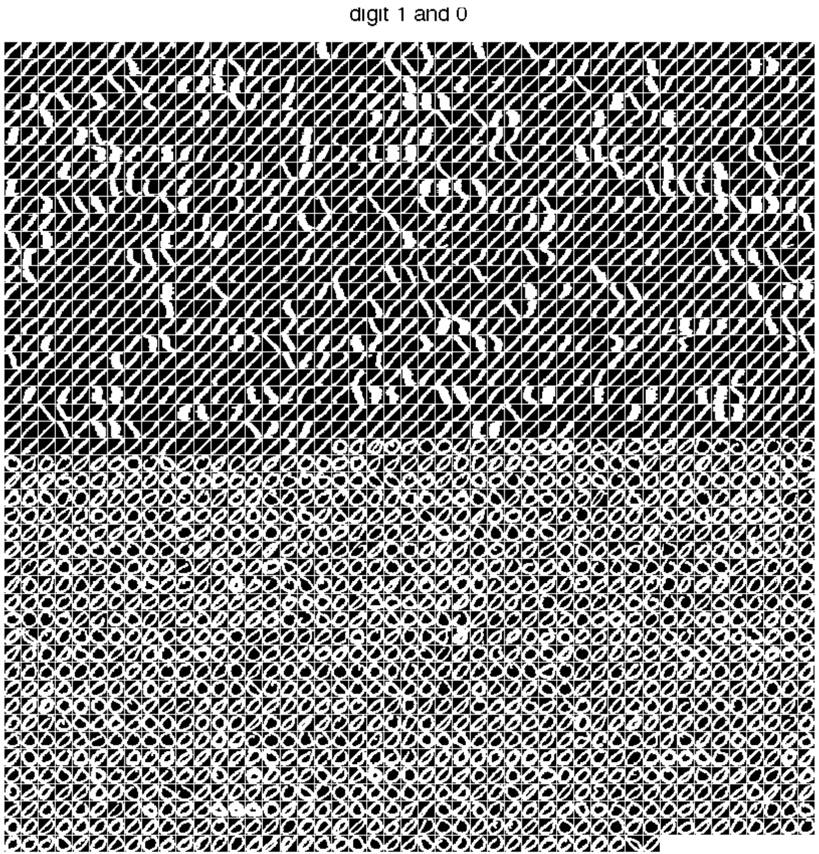
- Step 2: Take the eigenvectors $\boxed{w^1, w^2, \dots}$ of C corresponding to the largest eigenvalue λ_1, the second largest eigenvalue λ_2 …

- Step 3: Compute reduced representation (principle components)

$$z^i = \begin{pmatrix} {w^1}^{\top}(x^i - \mu)/\sqrt{\lambda_1} \\ {w^2}^{\top}(x^i - \mu)/\sqrt{\lambda_2} \\ \vdots \end{pmatrix}$$

Good youtube on calculating eigenvectors and eigenvalue
https://www.youtube.com/watch?v=cHOsd2PhkqE

Run demo PCA_digits.m



Run demo PCA_leaf.m

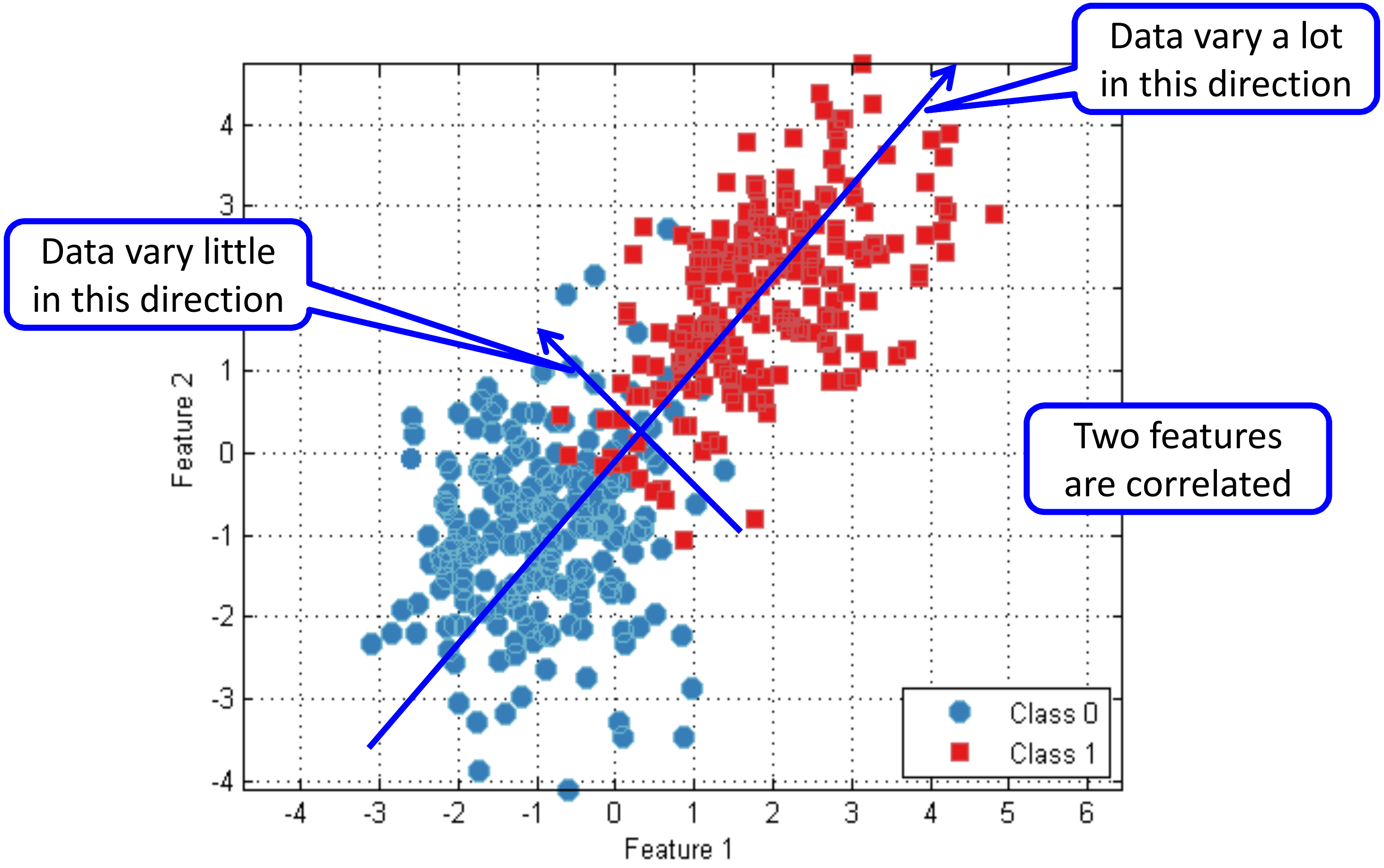
Use what criterion for reduction?

- There are many criteria (geometric based, information theory based, etc.)

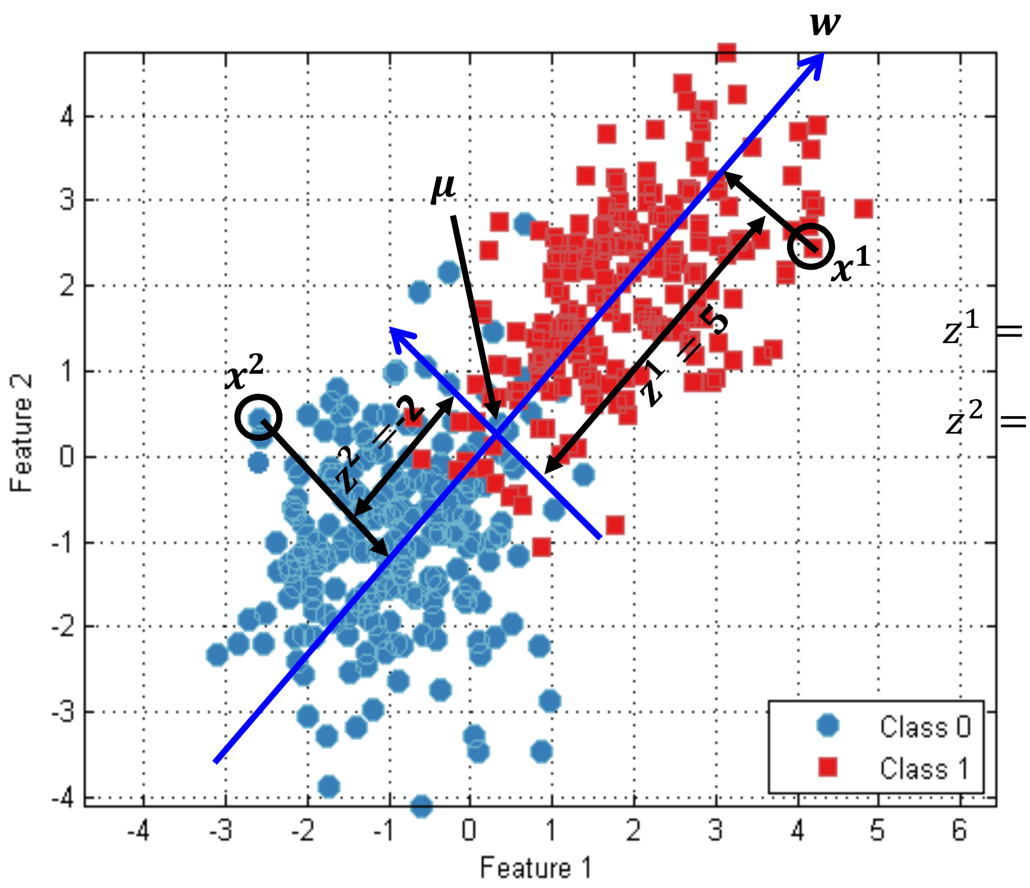
- One useful criterion: want to capture variation in data
  - variations are “signals” or information in the data
  - need to normalize each variables first

- In the process, also discover variables or dimensions highly correlated
  - represent highly related features
  - combine them to form a stronger signal
  - lead to simpler presentation

An example


Data vary a lot
in this direction
Data vary little
in this direction
Two features
are correlated
Feature 2
Feature 1
4
3
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
5
6
Class 0
Class 1

Reduce to 1 dimension



$$z^1 = w^\top (x^1 - \mu)$$

$$z^2 = w^\top (x^2 - \mu)$$

How to formulate the problem

- Given m data points, $\{x^1, x^2, \dots x^m\} \in R^n$, with their mean $\mu = \frac{1}{m}\sum_{i=1}^{m} x^i$

- Find a direction $w \in R^n$ where $\|w\| \le 1$

- Such that the variance (or variation) of the data along direction w is maximized

$$\max_{w:\|w\|\le 1} \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(w^\top x^i - w^\top \mu\right)^2}_{\text{variance}}$$

Is it an easy optimization problem?

- Manipulate the objective with linear algebra

$$\frac{1}{m}\sum_{i=1}^{m}\left(w^{\top}x^{i}-w^{\top}\mu\right)^{2}$$

$$=\frac{1}{m}\sum_{i=1}^{m}(w^{\top}(x^{i}-\mu))^{2}$$

$$=\frac{1}{m}\sum_{i=1}^{m}w^{\top}\left(x^{i}-\mu\right)\left(x^{i}-\mu\right)^{\top}w$$

$$=w^{\top}\underbrace{\left(\frac{1}{m}\sum_{i=1}^{m}\left(x^{i}-\mu\right)\left(x^{i}-\mu\right)^{\top}\right)}_{\text{covariance matrix } C}w$$

Landscape of the optimization problem

- Suppose the data has two dimension ($n = 2$)

- C is a diagonal matrix

$$C = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

- The optimization problem becomes

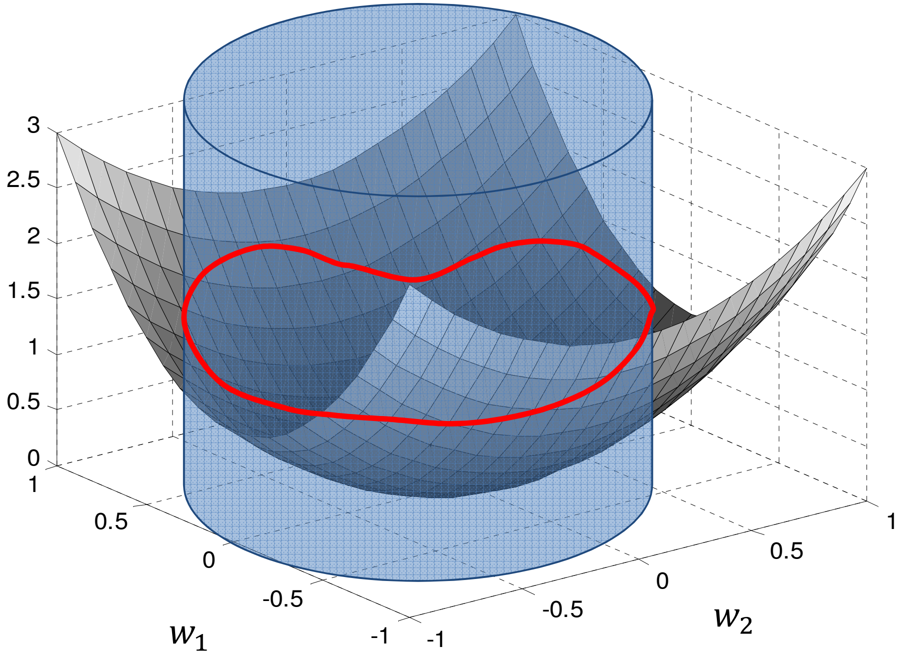
$$\max_{w:\|w\|\leq 1} w^\top C w$$

$$= \max_{w:\|w\|\leq 1} (w_1, w_2) \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

$$= \max_{w:\|w\|\leq 1} w_1^2 + 2w_2^2$$

Landscape of the optimization problem

- $f(w_1, w_2) = w_1^2 + 2w_2^2$



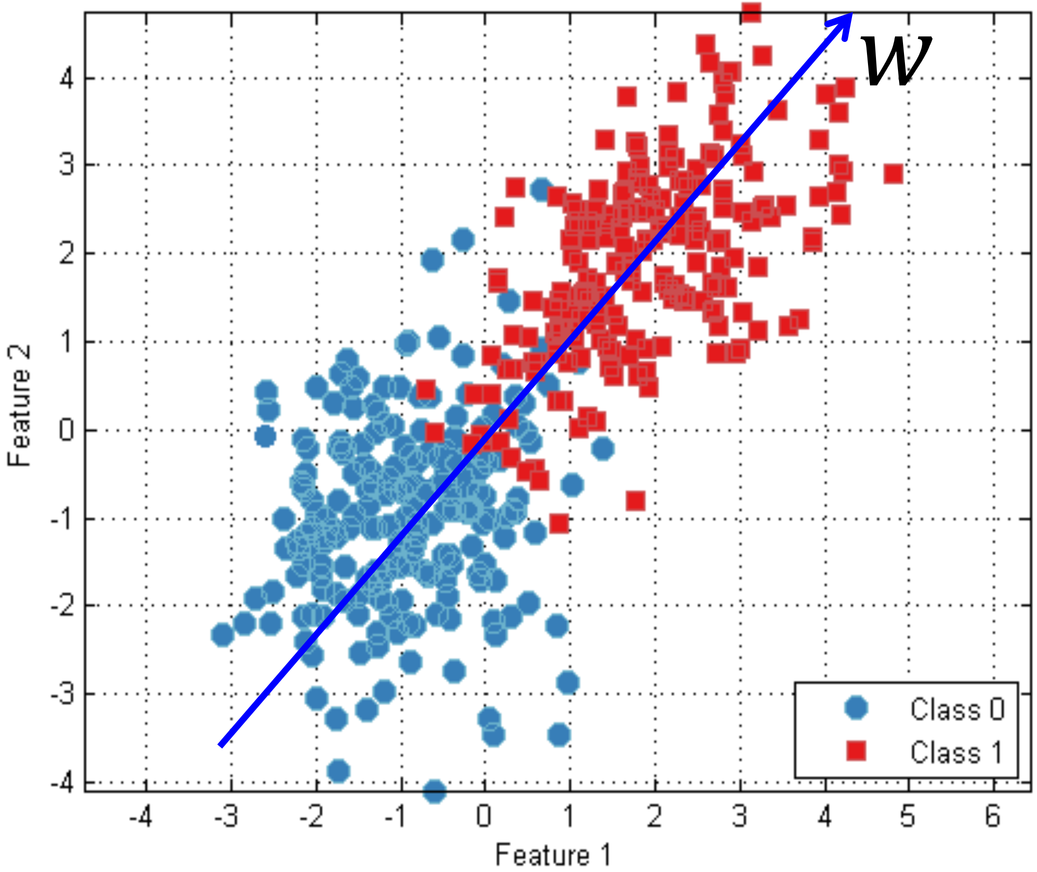
Eigen-value problem

- Eigen-value problem
  - Given a symmetric matrix $C \in R^{n \times n}$
  - Find a vector $w \in R^n$ and $\|w\| = 1$
  - Such that

$$Cw = \lambda w$$

- There will be multiple solution of $w^1, w^2, \ldots$ with different $\lambda_1, \lambda_2, \ldots$
  - They are ortho-normal: ${w^i}^\top w^i = 1, {w^i}^\top w^j = 0$

Principal direction of the data



Variance of in the principal direction

- Principal direction w satisfies

$$Cw = \lambda w$$

- Variance in principal direction is

$$w^\top C w$$

$$= \lambda w^\top w$$

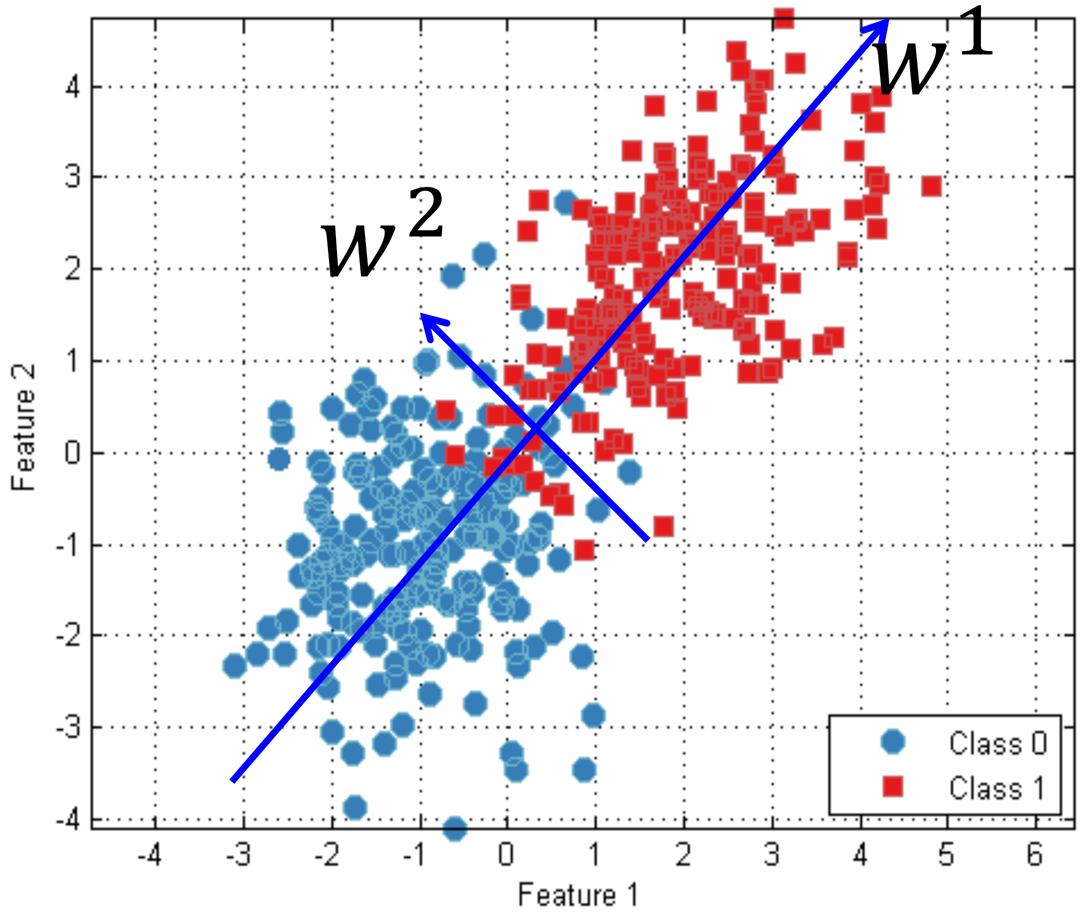
$$= \lambda$$

eigen-value

Find multiple principal directions

- Directions $w^1, w^2, \ldots$ which has
  - the largest variances
  - but are orthogonal to each other
- Take the eigenvectors $w^1, w^2, \ldots$ of C corresponding to
  - the largest eigenvalue λ_1,
  - the second largest eigenvalue λ_2
  - …

Principal direction of the data



Principal component analysis

- Given m data points, $\{x^1, x^2, \ldots x^m\} \in R^d$, with mean

- Step 1: Estimate the mean and covariance matrix from data

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^i \quad and \quad C = \frac{1}{m}\sum_{i=1}^{m}(x^i - \mu)(x^i - \mu)^\top$$

Principal directions

- Step 2: Take the eigenvectors $w^1, w^2, \ldots$ of C corresponding to the largest eigenvalue λ_1, the second largest eigenvalue λ_2 …

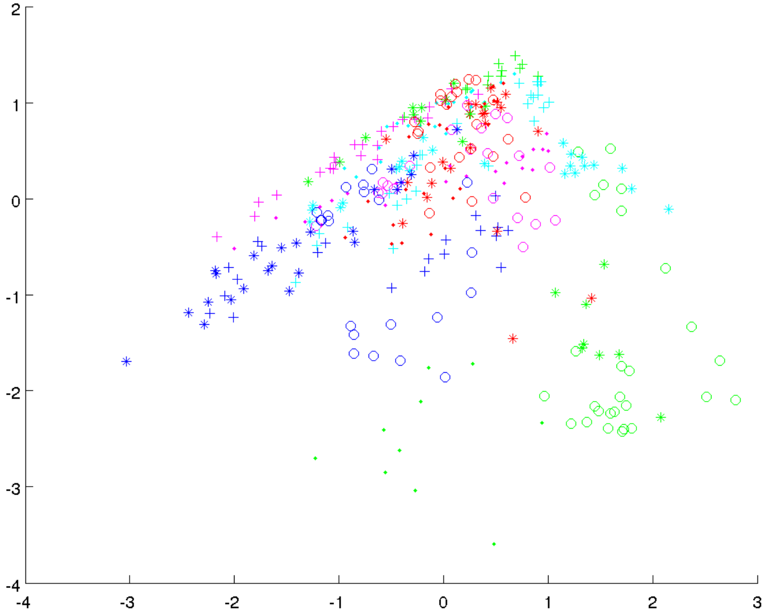
- Step 3: Compute reduced representation

$$z^i = \begin{pmatrix} {w^1}^\top (x^i - \mu)/\sqrt{\lambda_1} \\ {w^2}^\top (x^i - \mu)/\sqrt{\lambda_2} \\ \vdots \end{pmatrix}$$

Normalize by standard deviation

Look more into PCA_leaf.m

Interpreting the reduced representation



Principal direction:

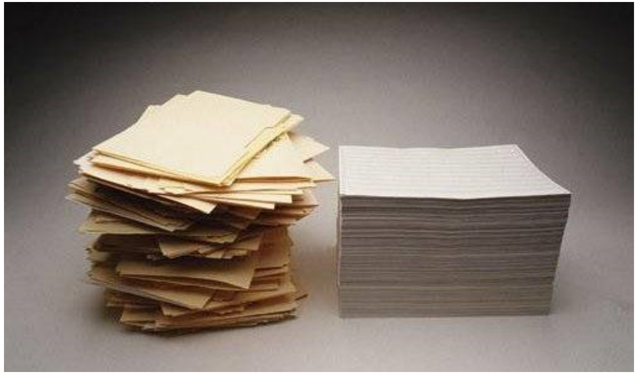
W =

0.0938	0.1924
0.1902	0.0253
0.2266	-0.1800
-0.1850	0.4084
-0.1600	0.3825
-0.2063	0.3488
0.1940	-0.4037
0.2150	-0.3566
-0.3723	-0.2001
-0.3657	-0.1974
-0.3602	-0.2037
-0.3175	-0.1886
-0.3056	-0.1243
-0.3482	-0.1829

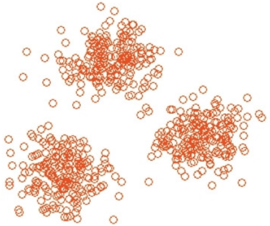
Shape features

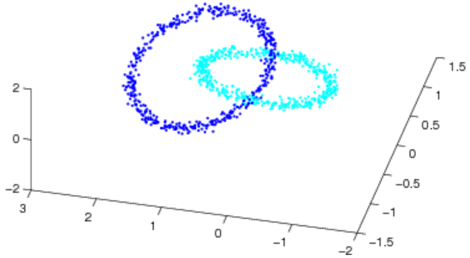
Texture features

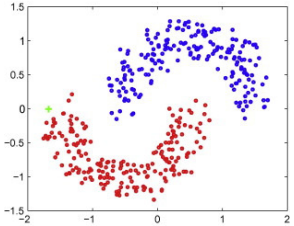
Documents collections



What are the relations between data points?

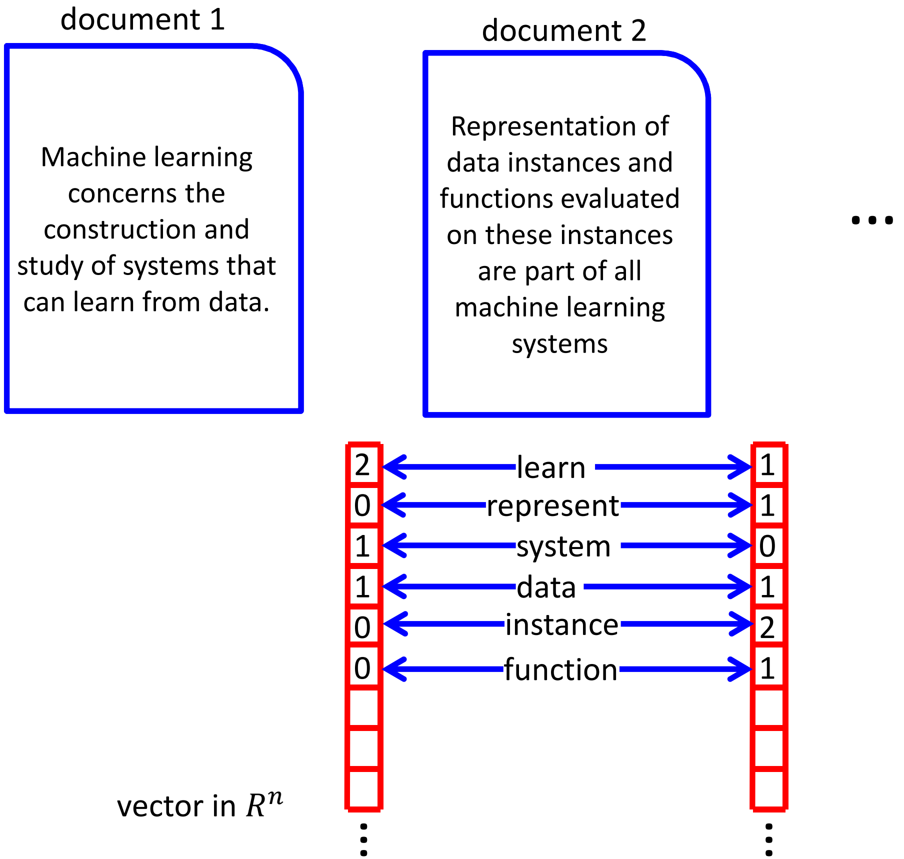






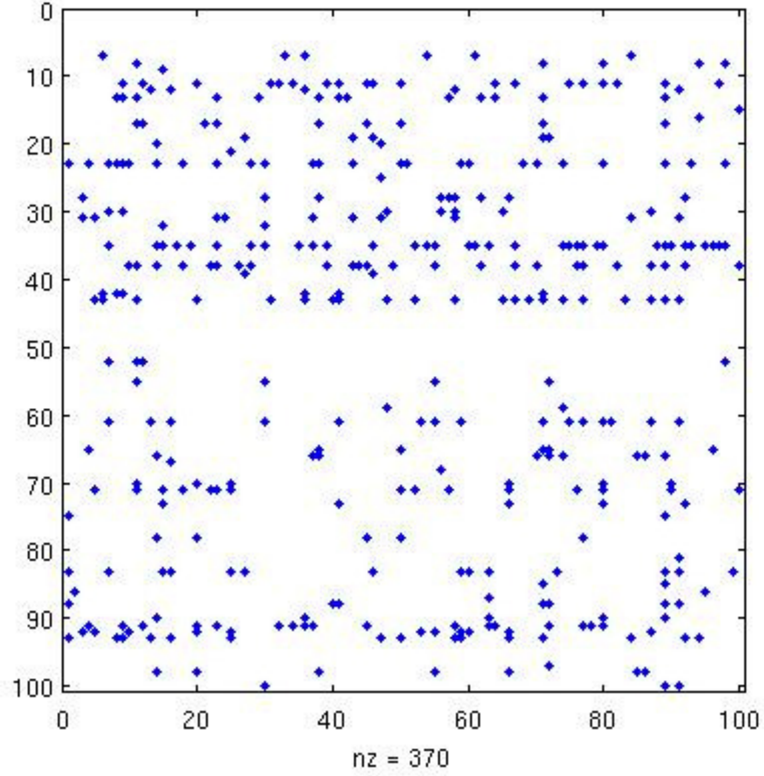
http://submitpapers.com/misc/docCollection.html

Bag of words representation



Experiments with 20 news groups

- Bag-of-words, or term-document matrix

$X =$ 

Singular Value Decomposition

- Singular value decomposition, known as SVD, is a factorization of a real matrix with applications in calculating pseudo-inverse, rank, solving linear equations, and many others.

- For a matrix $M \in \mathbb{R}^{m \times n}$ assume $n \leq m$
  - $M = U\Sigma V^\top$ where $U \in \mathbb{R}^{m \times m}$, $V^\top \in \mathbb{R}^{n \times n}$, $\Sigma \in \mathbb{R}^{m \times n}$

  - The m columns of U, and the n columns of V are called the left and right singular vectors of M. The diagonal elements of Σ, Σ_{ii} are known as the singular values of M.

  - Let v be the i^{th} column of V, and u be the i^{th} column of U, and σ be the i^{th} diagonal element of Σ

$$Mv = \sigma u \quad \text{and} \quad M^\top u = \sigma v$$

Singular Value Decomposition - II

- $M = [u_1 \quad u_2 \; \ldots \quad u_n] \begin{bmatrix} \Sigma_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_{mn} \end{bmatrix} [v_1 \quad v_2 \ldots \quad v_n]^\top$

  principal directions

  Scaling factor

  Projection in principal directions

- Singular value decomposition is related to eigenvalue decomposition
  - Suppose $M = [x_1 - \mu \quad x_2 - \mu \ldots \quad x_m - \mu] \in \mathbb{R}^{m \times n}$
  - Then covariance matrix is $C = \frac{1}{m} M M^\top$
  - Starting from singular vector pair
    - $M^\top u = \sigma v$
    - $\Rightarrow M M^\top u = \sigma M v$
    - $\Rightarrow M M^\top u = \sigma^2 u$
    - $\Rightarrow C u = \lambda u$


SUMMARY