Computational Data Analysis

Machine Learning

Yao Xie, Ph.D.

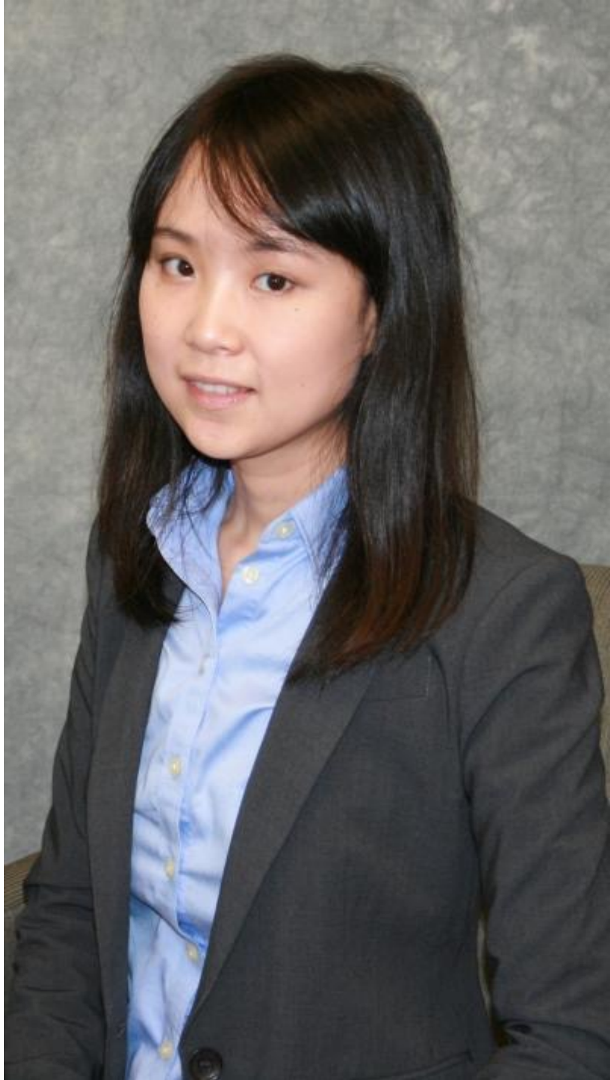
Associate Professor

Harold R. and Mary Anne Nash Early Career Professor

H. Milton Stewart School of Industrial and Systems Engineering

Dimensionality Reduction

Principal Component Analysis



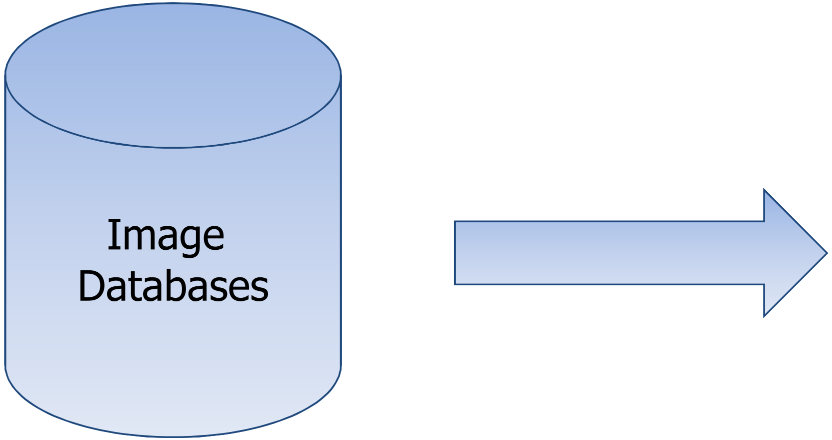
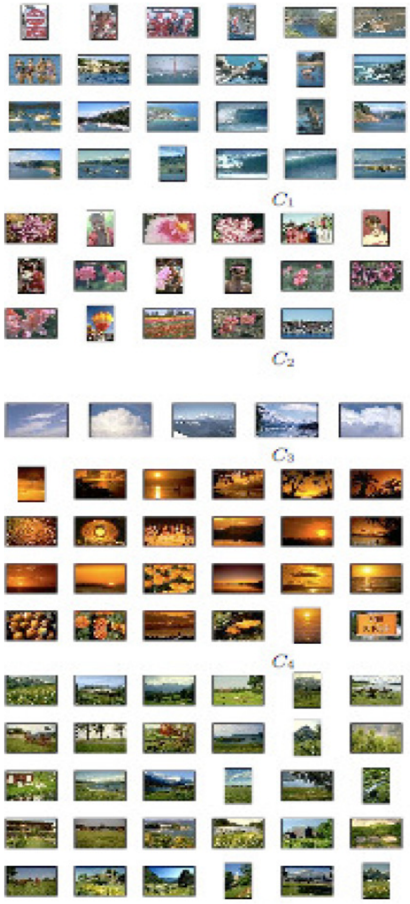
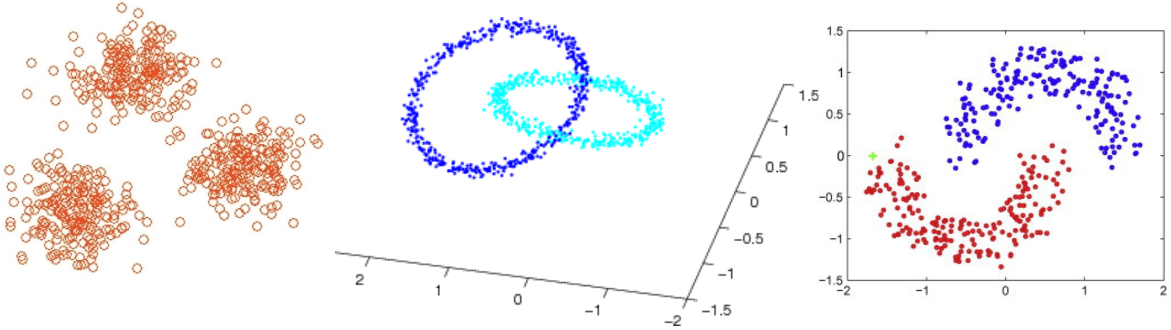


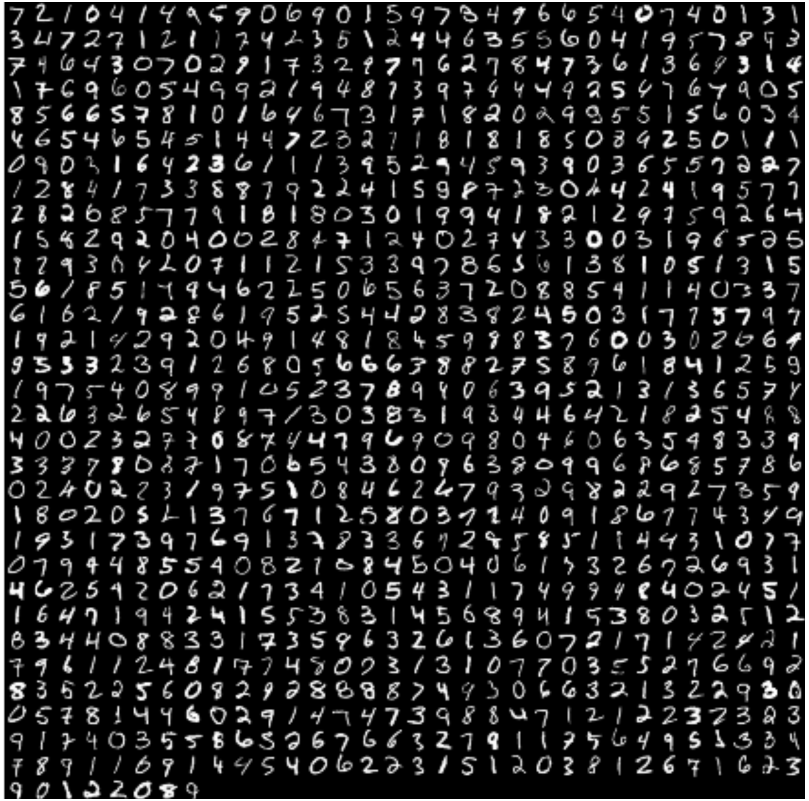
Image Databases



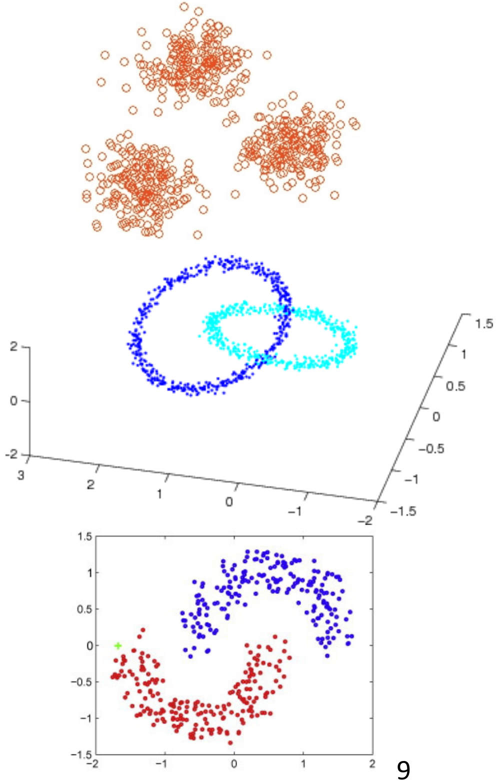
What are the relations between data points?



Handwritten digits



What are the relations between data points?

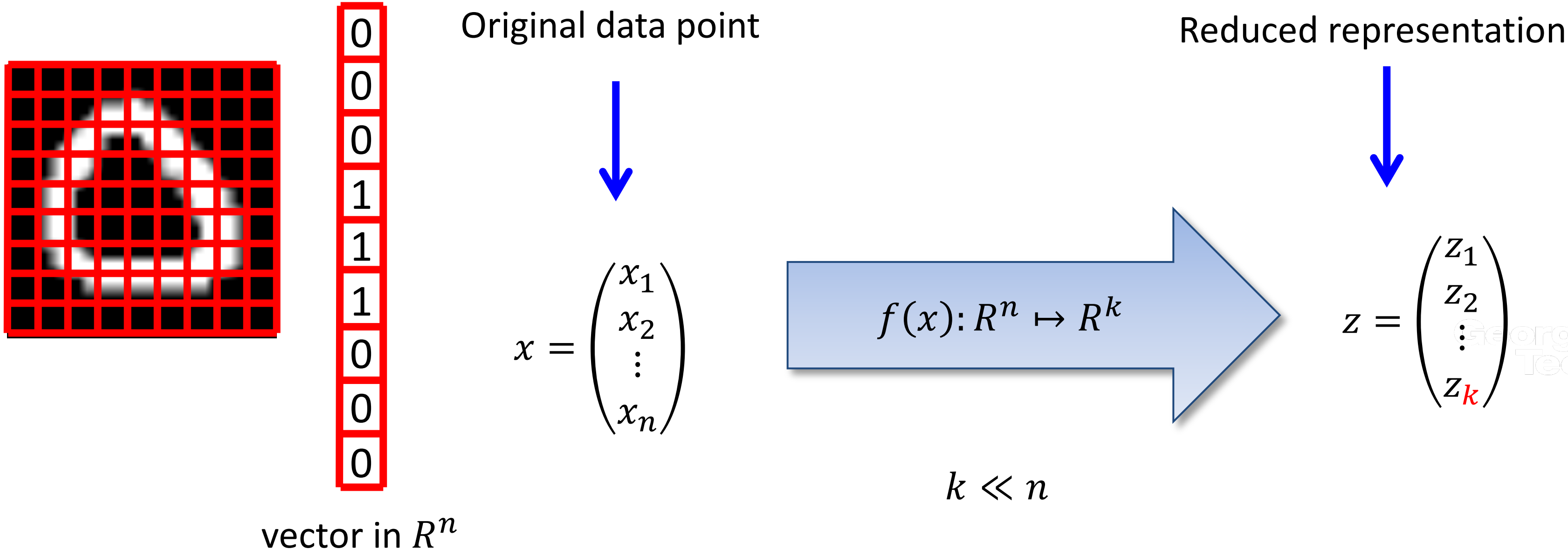


http://www.datawrangling.org/python-montage-code-for-displaying-arrays/

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



vector in R^n

Original data point

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$f(x): R^n \mapsto R^k$$

$$k \ll n$$

Reduced representation

$$z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{pmatrix}$$

Why dimensionality reduction and how to think

- The dimension-reduced data can be used for
  - Visualizing, exploring and understanding the data
  - Extracting ”features” and dominant modes
  - Cleaning data
  - Speeding up subsequent learning task
  - Building simpler model later

- Applications
  - Image compression
  - Face recognition (eigenface)
  - Natural language processing (latent semantic analysis)

Principal component analysis

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

Step 1: Estimate the mean and covariance matrix from data

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

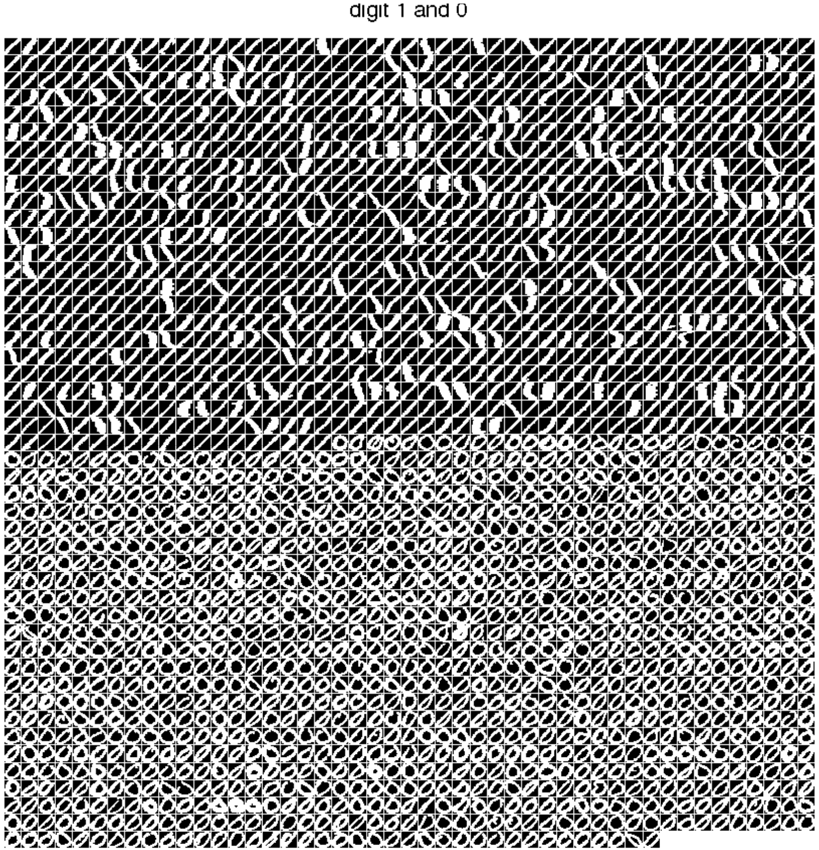
Weight vectors

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

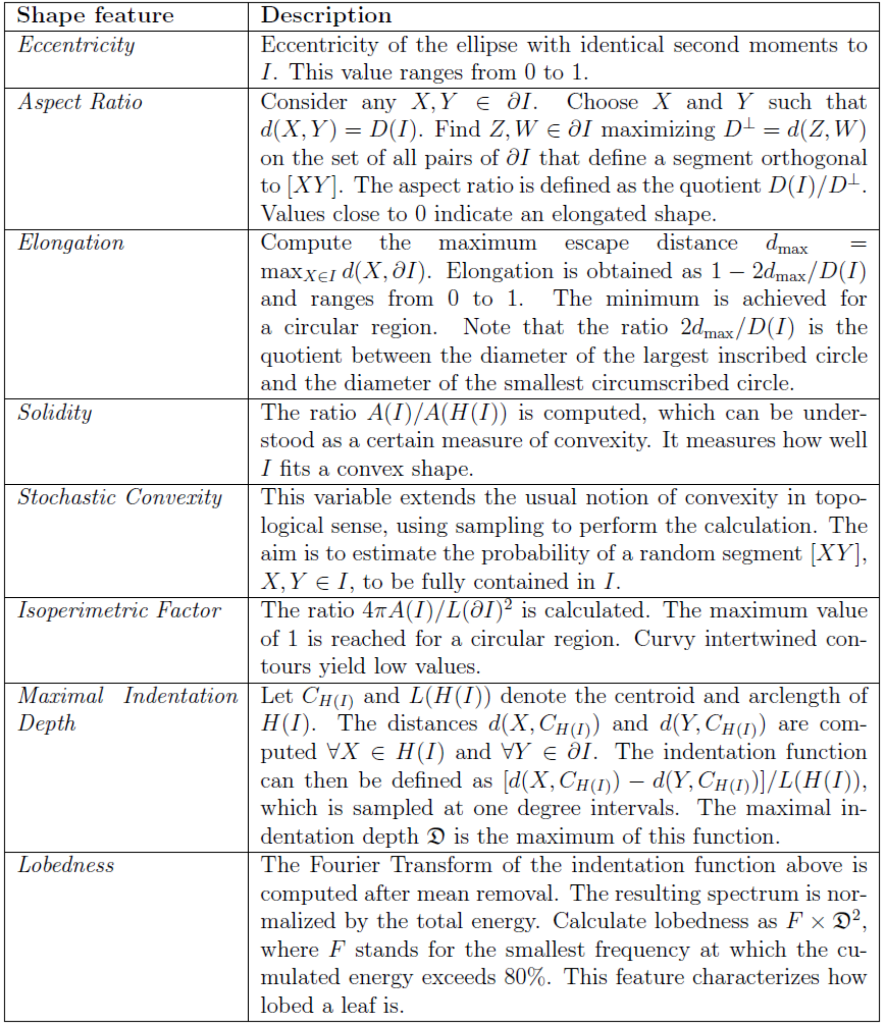
Step 3: Compute reduced representation (**principle components** of a data point)

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

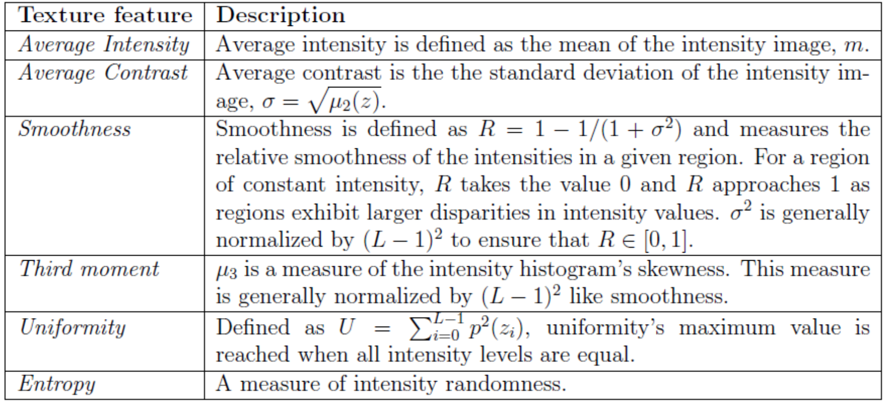
Run demo PCA_digits.m



Run demo PCA_leaf.m



Shape feature	**Description**
Eccentricity	Eccentricity of the ellipse with identical second moments to I. This value ranges from 0 to 1.
Aspect Ratio	Consider any $X, Y \in \partial I$. Choose X and Y such that $d(X,Y) = D(I)$. Find $Z, W \in \partial I$ maximizing $D^{\perp} = d(Z,W)$ on the set of all pairs of ∂I that define a segment orthogonal to $[XY]$. The aspect ratio is defined as the quotient $D(I)/D^{\perp}$. Values close to 0 indicate an elongated shape.
Elongation	Compute the maximum escape distance $d_{\max} = \max_{X \in I} d(X, \partial I)$. Elongation is obtained as $1 - 2d_{\max}/D(I)$ and ranges from 0 to 1. The minimum is achieved for a circular region. Note that the ratio $2d_{\max}/D(I)$ is the quotient between the diameter of the largest inscribed circle and the diameter of the smallest circumscribed circle.
Solidity	The ratio $A(I)/A(H(I))$ is computed, which can be understood as a certain measure of convexity. It measures how well I fits a convex shape.
Stochastic Convexity	This variable extends the usual notion of convexity in topological sense, using sampling to perform the calculation. The aim is to estimate the probability of a random segment $[XY]$, $X, Y \in I$, to be fully contained in I.
Isoperimetric Factor	The ratio $4\pi A(I)/L(\partial I)^2$ is calculated. The maximum value of 1 is reached for a circular region. Curvy intertwined contours yield low values.
Maximal Indentation Depth	Let $C_{H(I)}$ and $L(H(I))$ denote the centroid and arclength of $H(I)$. The distances $d(X, C_{H(I)})$ and $d(Y, C_{H(I)})$ are computed $\forall X \in H(I)$ and $\forall Y \in \partial I$. The indentation function can then be defined as $[d(X, C_{H(I)}) - d(Y, C_{H(I)})]/L(H(I))$, which is sampled at one degree intervals. The maximal indentation depth $\mathfrak{D}$ is the maximum of this function.
Lobedness	The Fourier Transform of the indentation function above is computed after mean removal. The resulting spectrum is normalized by the total energy. Calculate lobedness as $F \times \mathfrak{D}^2$, where F stands for the smallest frequency at which the cumulated energy exceeds 80%. This feature characterizes how lobed a leaf is.



Texture feature	**Description**
Average Intensity	Average intensity is defined as the mean of the intensity image, m.
Average Contrast	Average contrast is the the standard deviation of the intensity image, $\sigma = \sqrt{\mu_2(z)}$.
Smoothness	Smoothness is defined as $R = 1 - 1/(1 + \sigma^2)$ and measures the relative smoothness of the intensities in a given region. For a region of constant intensity, R takes the value 0 and R approaches 1 as regions exhibit larger disparities in intensity values. σ^2 is generally normalized by $(L-1)^2$ to ensure that $R \in [0, 1]$.
Third moment	μ_3 is a measure of the intensity histogram's skewness. This measure is generally normalized by $(L-1)^2$ like smoothness.
Uniformity	Defined as $U = \sum_{i=0}^{L-1} p^2(z_i)$, uniformity's maximum value is reached when all intensity levels are equal.
Entropy	A measure of intensity randomness.

8 shape features and
6 texture features

Use what criterion for reduction?

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

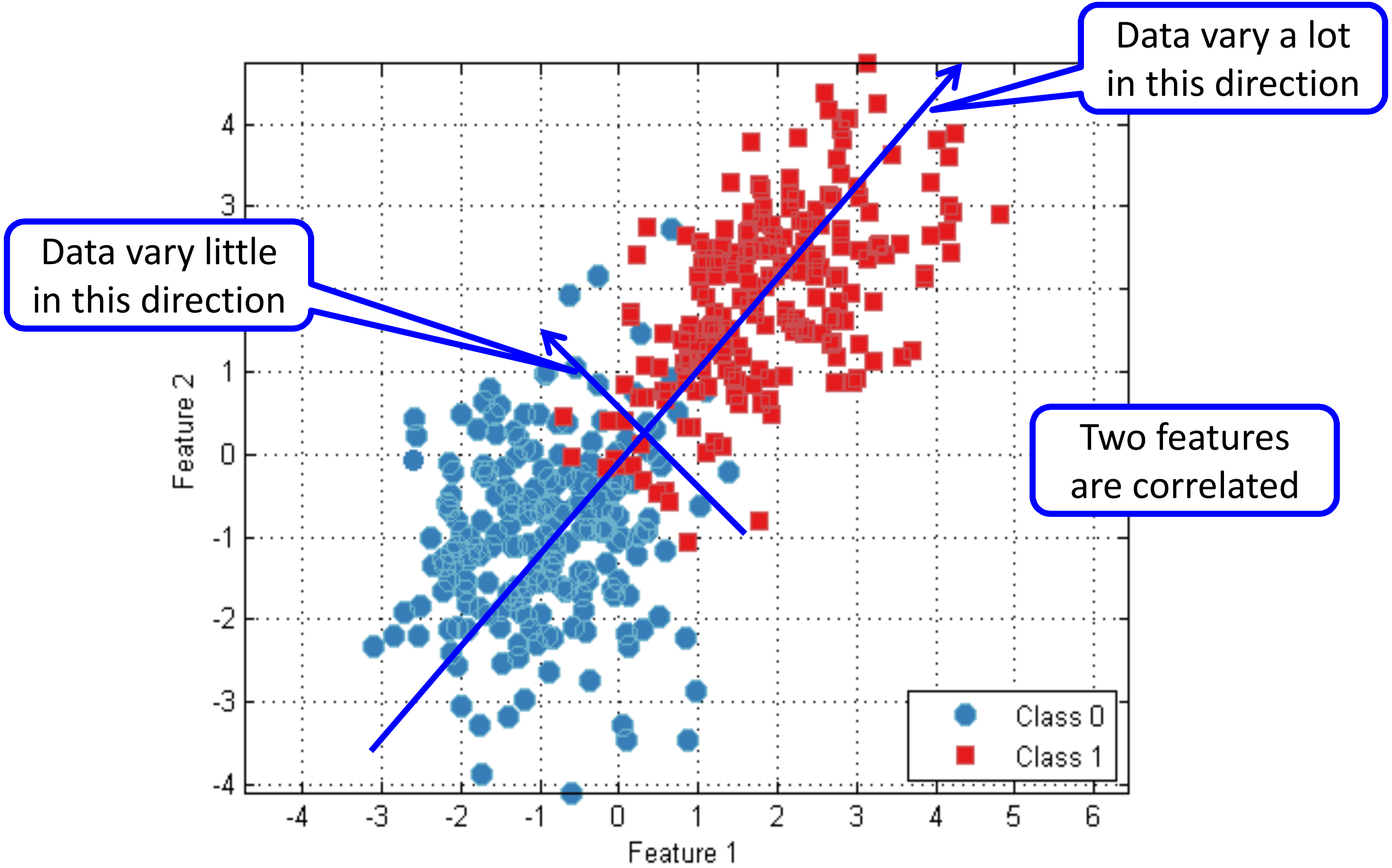
One criterion: want to capture variation in data

- 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 phenomena
- combine them to form a stronger signal
- lead to simpler presentation

An example

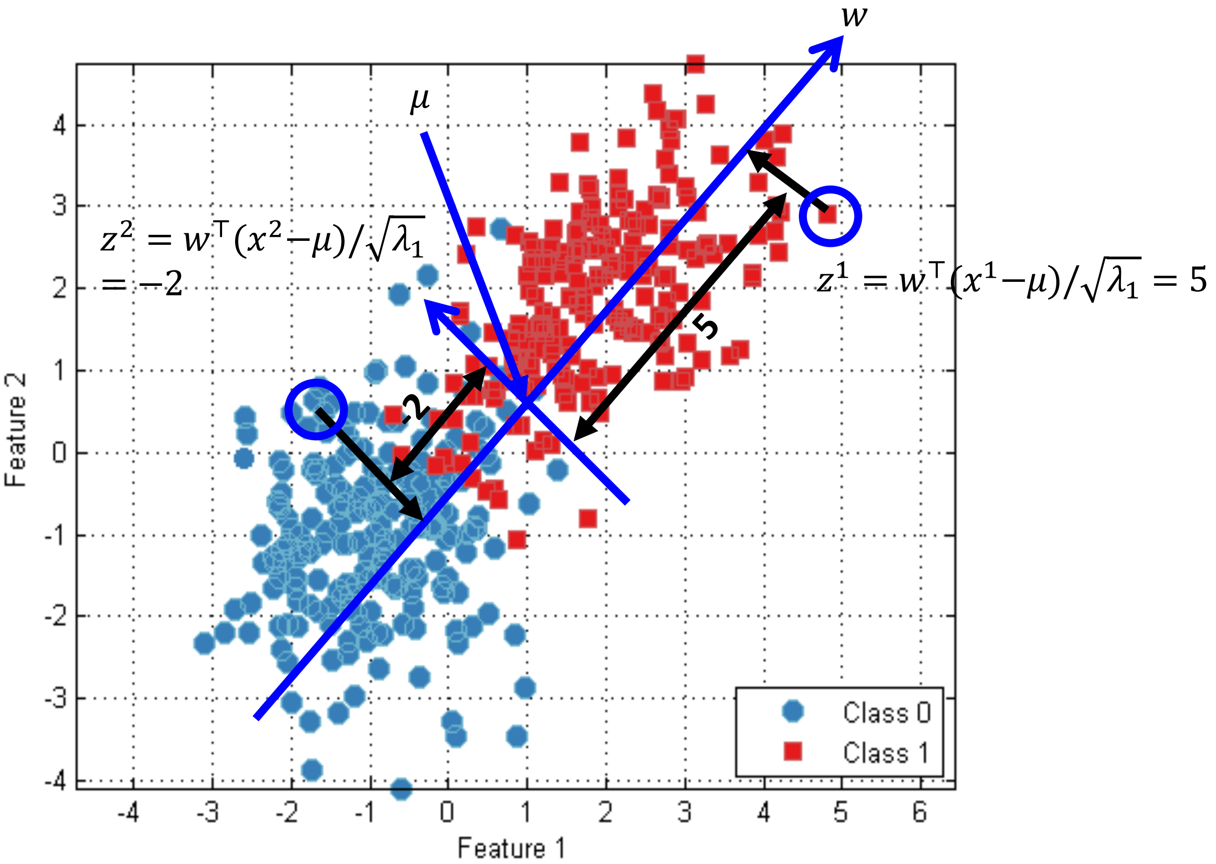


Data vary a lot in this direction

Data vary little in this direction

Two features are correlated

An example (cont.)



$z^2 = w^\top(x^2 - \mu)/\sqrt{\lambda_1}$
$= -2$

$z^1 = w^\top(x^1 - \mu)/\sqrt{\lambda_1} = 5$

How to formulate the problem

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

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

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

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

Is it an easy optimization problem?

Manipulate the objective with linear algebra

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\left(w^\top x^i - w^\top \mu\right)^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(w^\top (x^i - \mu)\right)^2 \\
&= \frac{1}{m}\sum_{i=1}^{m} w^\top (x^i - \mu)(x^i - \mu)^\top w \\
&= w^\top \underbrace{\left(\frac{1}{m}\sum_{i=1}^{m}(x^i - \mu)(x^i - \mu)^\top\right)}_{\text{covariance matrix } C} w
\end{aligned}$$

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

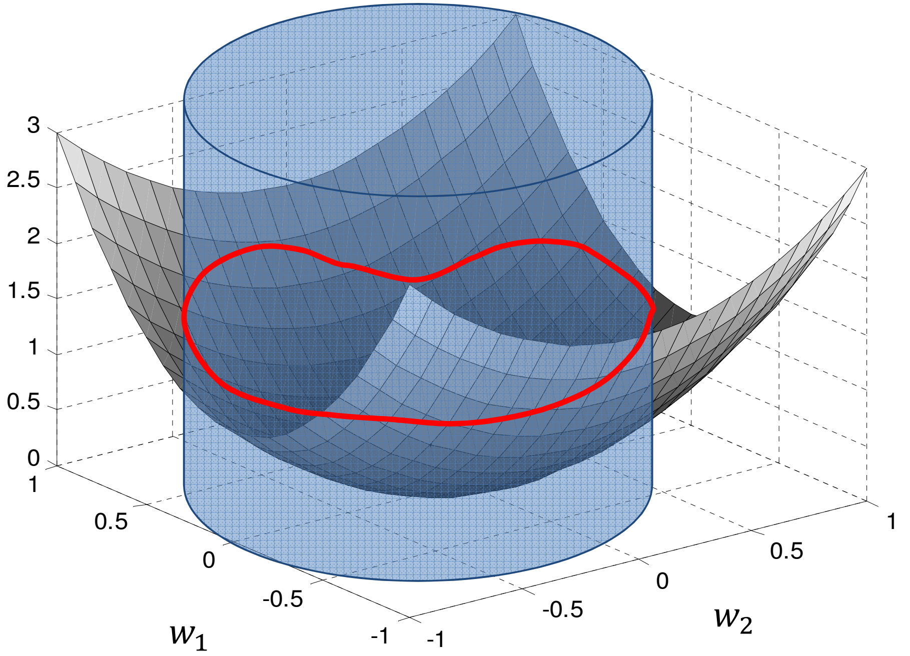
$$\max_{w:\|w\| \leq 1} w^\top C w$$

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

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

Landscape of the optimization problem

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



Solving the PCA problem

$$\max_{w:\|w\|\leq 1} w^\top C w, \quad C = \frac{1}{m}\sum_{i=1}^{m}(x^i - \mu)(x^i - \mu)^\top$$

- Form Lagrangian function of the optimization problem

$$L(w,\lambda) = w^\top C w + \lambda(1 - \|w\|^2)$$

- If w is a maximum of the original optimization problem, then there exists a λ, where (w,λ) is a stationary point of $L(w,\lambda)$
- This implies that

$$\frac{\partial L}{\partial w} = 0 = 2Cw - 2\lambda w \Leftrightarrow Cw = \lambda w$$

- The optimal solution w should be an eigen-vector of C
- Objective function becomes λ (associated with w)

Variance of in the principal direction

- The optimal solution w should be an eigen-vector of C

$$Cw = \lambda w$$

- Objective function becomes λ (associated with w)

$$w^\top C w = \lambda w^\top w = \lambda||w||^2$$

eigen-value

- The problem becomes finding the largest eigenvalue of C

Eigenvalue problem

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

$$Cu = \lambda u$$

- There will be multiple solution: $u^1, u^2, \dots u^n$ (called the **eigenvectors**) with different $\lambda_1, \lambda_2, \dots \lambda_n$ (called the **eigenvalues**.)
  - Eigenvectors are ortho-normal: ${u^i}^\top u^i = 1, {u^i}^\top u^j = 0$
  - Eigenvalues are called spectrum
- Eigendecomposition

$$C = U \Lambda U^T$$

Find multiple principal directions

Directions $w^1, w^2, \ldots$ which has

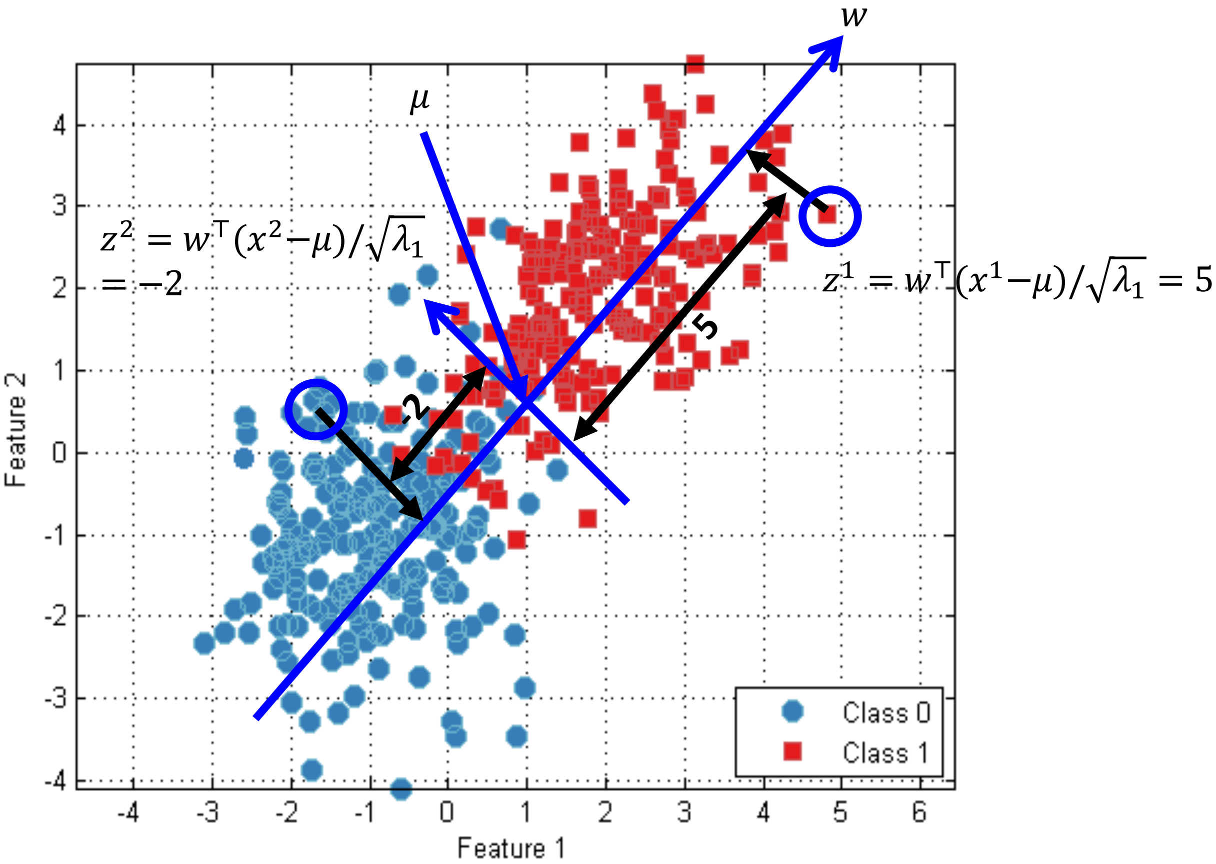
- the largest variances
- orthogonal to each other

Take the eigenvectors $w^1, w^2, \ldots$ of C corresponding to

- the largest eigenvalue λ_1,
- the second largest eigenvalue λ_2

and so on.

An example (cont.)



Principal component analysis

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

Step 1: Estimate the mean and covariance matrix from data

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

Weight vectors

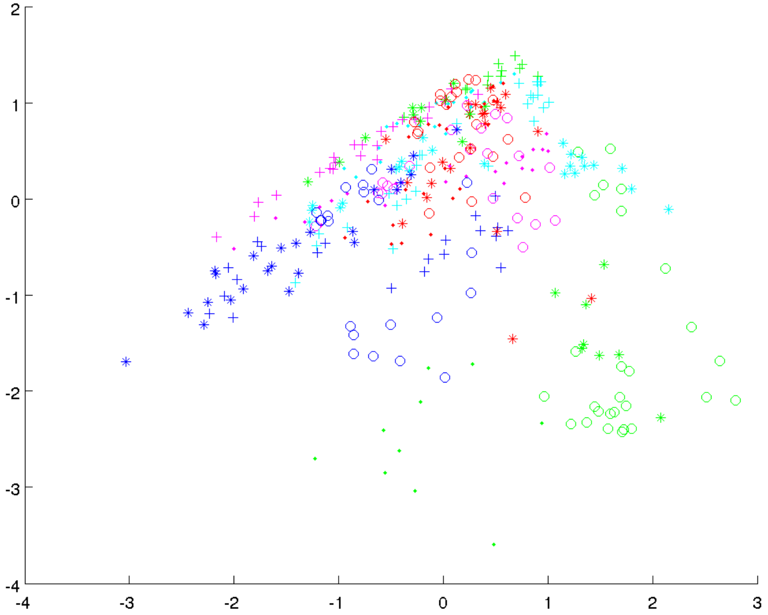
Step 2: Take the eigenvectors $w^1, w^2, \dots$ of C corresponding to the largest eigenvalue λ_1, the second largest eigenvalue λ_2 …

Step 3: Compute reduced representation (**principle components** of a data point)

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

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

Singular Value Decomposition (SVD)

- Singular value decomposition, known as SVD, is a factorization of a real matrix
- For a matrix $M \in \mathbb{R}^{n \times m}$ $(n \leq m)$

$$M = U \Sigma V^{\top}$$

$$M = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & \sigma_n & \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \dots & v_m \end{bmatrix}^{\top}$$

$U \in R^{n \times n}$
Left singular vectors
(orthonormal)

$\Sigma \in R^{n \times m}$
Singular values

$V \in R^{m \times m}$
Right singular vectors
(orthonormal)

where $U \in \mathbb{R}^{n \times n}$, $V^{\top} \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{n \times m}$

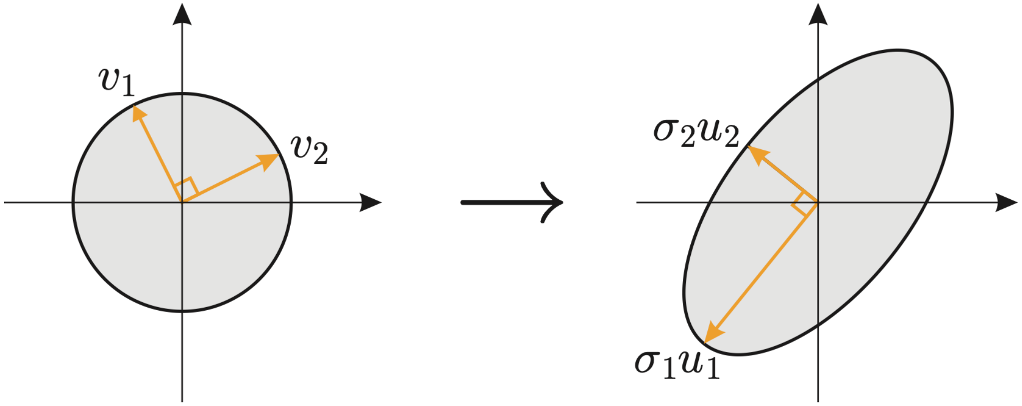
Typically $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$

Interpretations of SVD

- A pair of singular vectors (u, v) satisfies

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

- Geometry



http://ee263.stanford.edu/lectures/svd-v2.pdf

Relationship between SVD and eigendecomposition

$$M = U\Sigma V^{\top}$$

$$C := MM^T = U\Sigma V^T V\Sigma^T U^T = U\boxed{\Sigma\Sigma^T}U^T$$

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n^2 \end{bmatrix}$$

- The eigenvectors of $C := MM^T$ is U (the left singular vectors of M)
- The eigenvalues of C is σ_i^2 (squared singular values of M)
- Similar results can be derived for $M^{\mathrm{T}}M$

Another way to perform PCA (using SVD)

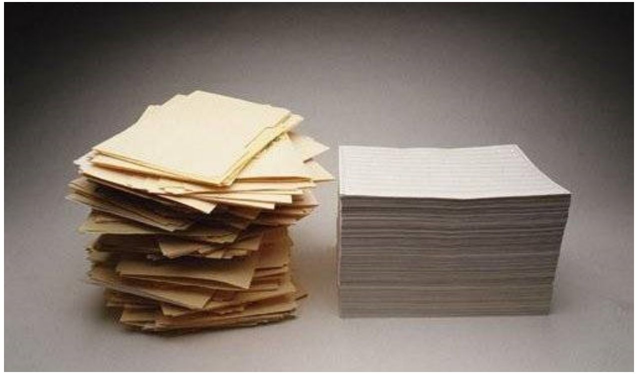
- Note that data covariance matrix can be written as

$$C = \frac{1}{m}\sum_{i=1}^{m}\left(x^i - \mu\right)\left(x^i - \mu\right)^{\mathsf{T}} = \frac{1}{m}XX^T$$

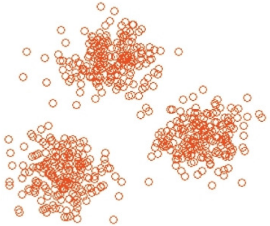
$$X = [x^1 - \mu, \cdots, x^m - \mu] \in R^{n\times m}$$

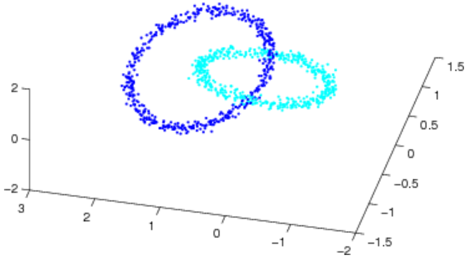
- Eigenvectors of C corresponds to left singular vectors of X
- Find the weight vectors $\{w^1, w^2, \cdots, w^r\}$ as the r left singular vectors of the data matrix X (r is the number of principle components)

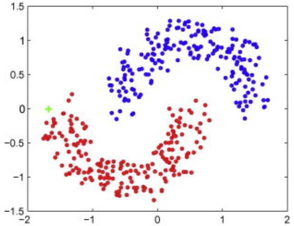
Documents collections



What are the relations between data points?

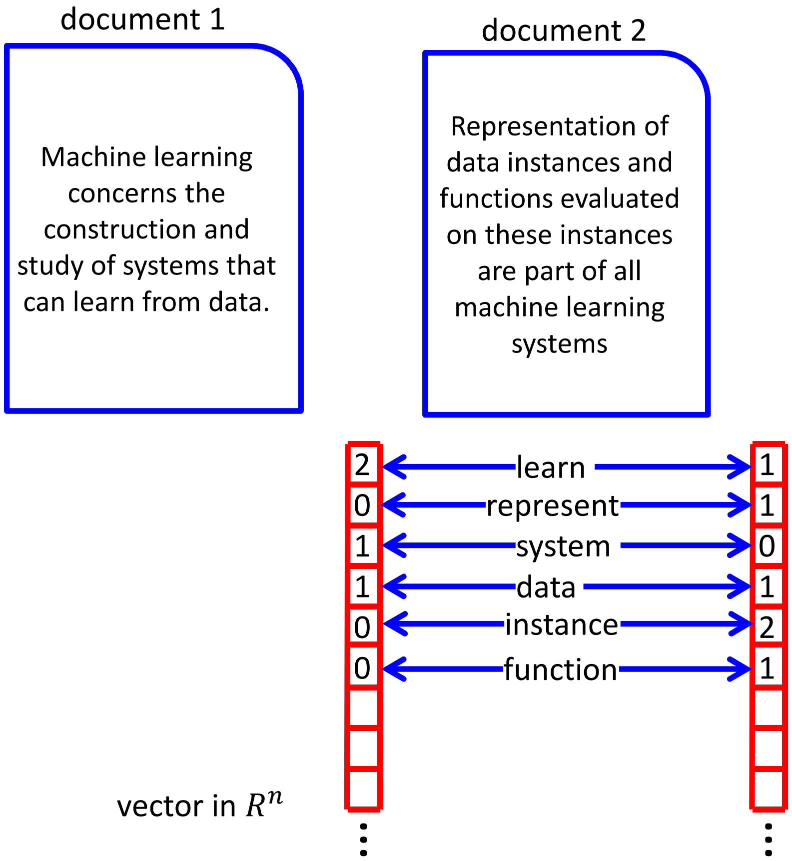






http://submitpapers.com/misc/docCollection.html

Bag of words representation



document 1

Machine learning concerns the construction and study of systems that can learn from data.

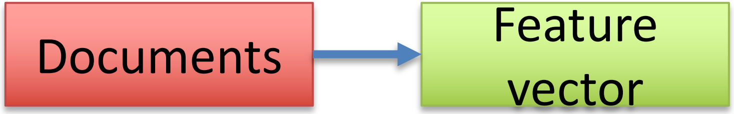
document 2

Representation of data instances and functions evaluated on these instances are part of all machine learning systems

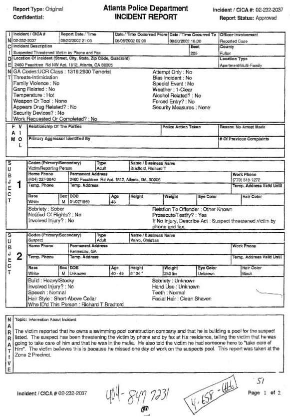
2	learn	1
0	represent	1
1	system	0
1	data	1
0	instance	2
0	function	1
⋮		⋮

vector in R^n

…



Documents → Feature vector



Report Type: Original
Confidential:

Atlanta Police Department
INCIDENT REPORT

Incident / CICA #: 02-232-2037
Report Status: Approved

INCIDENT					
	Incident / CICA # 02-232-2037	Report Date / Time 08/20/2002 21:03	Date / Time Occurred From 08/06/2002 09:00	Date / Time Occurred To 08/20/2002 18:00	Officer Involvement Reported Case
	Incident Description Suspected Threatened Victim by Phone and Fax			Beat 205	County Fulton
	Location Of Incident (Street, City, State, Zip Code, Quadrant) 2460 Peachtree Rd NW Apt. 1812, Atlanta, GA 30305				Location Type Apartment/Multi-Family

GA Codes:UCR Class : 1316:2600 Terrorist Threats-Intimidation
Family Violence : No
Gang Related : No
Temperature : Hot
Weapon Or Tool : None
Appears Drug Related? : No
Security Devices? : No
Work Requested Or Completed? : No
Attempt Only : No
Bias Incident : No
Special Event : No
Weather : 1-Clear
Alcohol Related? : No
Forced Entry? : No
Security Measures : None

FAM VIOL			
	Relationship Of The Parties	Police Action Taken	Reason No Arrest Made
	Primary Aggressor Identified By		# Of Previous Complaints

SUBJECT 1							
Codes (Primary/Secondary) Victim/Reporting Person		Type Adult	Name / Business Name Bradford, Richard T				
Home Phone (404) 237-3840	Permanent Address 2460 Peachtree Rd Apt. 1812, Atlanta, GA, 30305						Work Phone (770) 318-1272
Temp. Phone	Temp. Address						Temp. Address Valid Until
Race White	Sex M	DOB 01/07/1959	Age 43	Height	Weight	Eye Color	Hair Color

Sobriety : Sober
Notified Of Rights? : No
Involved Injury? : No
Relation To Offender : Other Known
Prosecute/Testify? : Yes
If No Injury, Describe Act : Suspect threatened.victim by phone and fax.

SUBJECT 2							
Codes (Primary/Secondary) Suspect		Type Adult	Name / Business Name Valvo, Christian				
Home Phone	Permanent Address Kennesaw, GA						Work Phone
Temp. Phone	Temp. Address						Temp. Address Valid Until
Race White	Sex M	DOB Unknown	Age 40 - 45	Height 6 ' 04 "	Weight 240 lbs	Eye Color Unknown	Hair Color Black

Build : Heavy/Stocky
Involved Injury? : No
Speech : Normal
Hair Style : Short-Above Collar
Who ID'd This Person : Richard T Bradrord
Sobriety : Unknown
Hand Use : Unknown
Teeth : Normal
Facial Hair : Clean Shaven

NARRATIVE

Topic: Information About Incident

The victim reported that he owns a swimming pool construction company and that he is building a pool for the suspect listed. The suspect has been threatening the victim by phone and by fax at his residence, telling the victim that he was going to take care of him and that he was in the mafia. He also told the victim he had someone here to "take care of him". The victim believes this is because he missed one day of work on the suspects pool. This report was taken at the Zone 2 Precinct.

Incident / CICA # 02-232-2037

404- 847 7231

4-658-4446

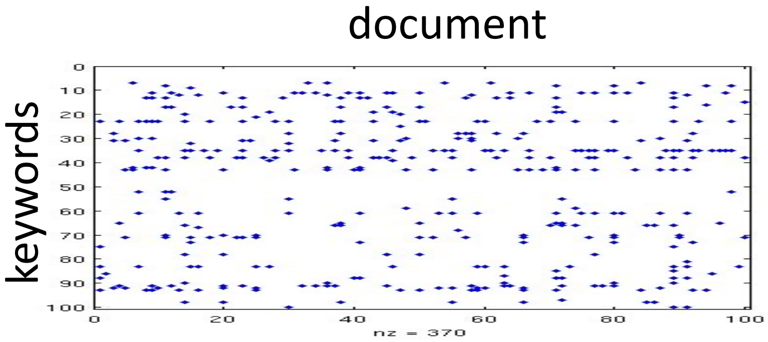
51

Page 1 of 2

Latent semantic analysis (LSA)

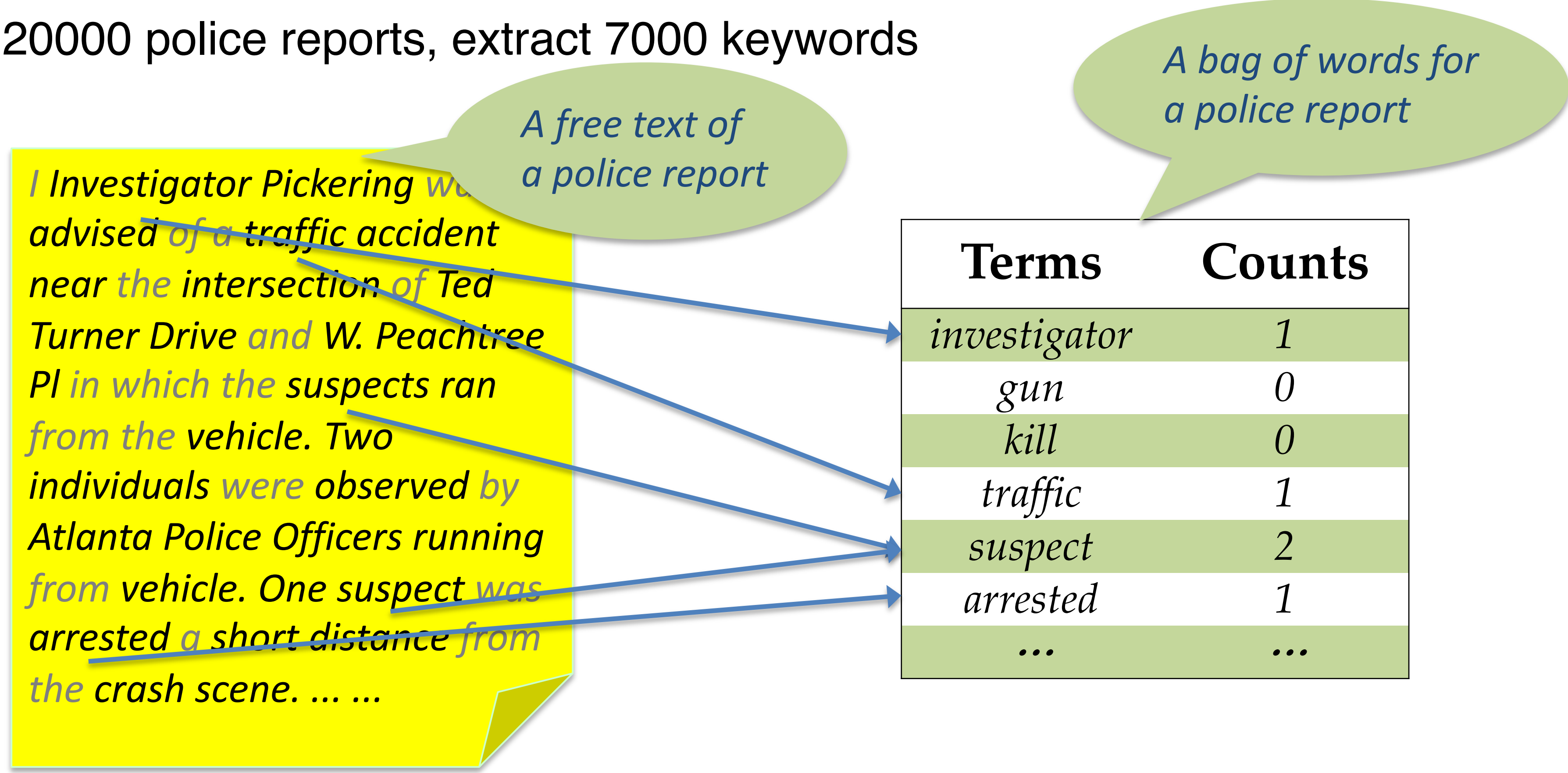
- Bag-of-words model or term-document matrix M (more natural language processing techniques: WF/IDF, removing stop words, N-gram)
- Perform PCA of M: (Latent Semantic Analysis, LSA)
- principle components z^i for each document can be interpreted as "feature" of a document

$X =$



Example: Atlanta Police reports

- 20000 police reports, extract 7000 keywords



A free text of a police report

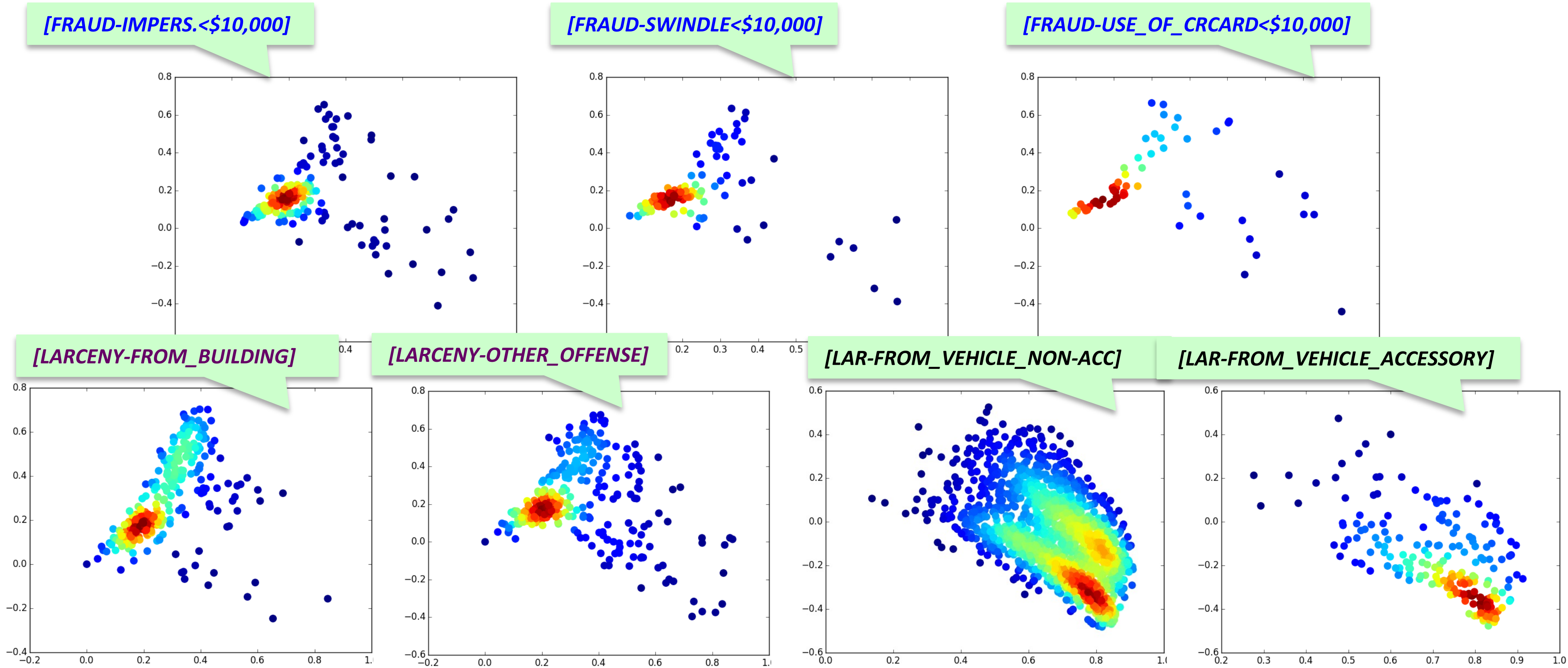
I Investigator Pickering wa advised of a traffic accident near the intersection of Ted Turner Drive and W. Peachtree Pl in which the suspects ran from the vehicle. Two individuals were observed by Atlanta Police Officers running from vehicle. One suspect was arrested a short distance from the crash scene. ... ...

A bag of words for a police report

Terms	Counts
investigator	*1*
gun	*0*
kill	*0*
traffic	*1*
suspect	*2*
arrested	*1*
…	…

Example: using PCA for data visualization

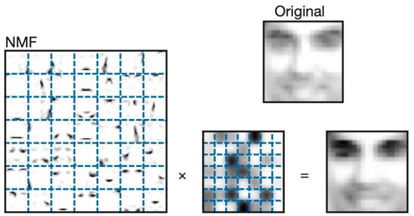
Atlanta police data, 20000 police reports, 7200 keywords (bi-gram), map into 2 principle components; shown 2d density estimation.



Extensions/variants of PCA

- Robust PCA: PCA is robust to outliers; it is a common practice to remove outliers to perform PCA; can be cast as a convex optimization problem (Candes, Li, Ma, Wright, 2009)
- Sparse PCA: traditional PCA combines all variables (using the weight vector), not ideal for high-dimensional data; sparse PCA combines just a few important features (solved by optimization) (Johnstone, Yu, 2009)
- Nonlinear PCA: kernelized PCA
- Nonnegative matrix factorization (NMF)
- ICA (independent component analysis):

decompose the signal into additive components



Lee, Seung, 1999, Nature.


SUMMARY