Dimensionality reduction

PCA, SVD, ICA

Slides taken from Jure Leskovec,
Christos Faloutsos and Tom Mitchell

Why dimensionality reduction?

- Some features may be irrelevant
- We want to visualize high dimensional data
- “Intrinsic” dimensionality may be smaller than the number of features

Supervised feature selection

- Scoring features:
  - Mutual information between attribute and class
  - χ^2: independence between attribute and class
  - Classification accuracy
- Domain specific criteria:
  - E.g. Text:
    - remove stop-words (and, a, the, …)
    - Stemming (going → go, Tom's → Tom, …)
    - Document frequency

Choosing sets of features

- Score each feature
- Forward/Backward elimination
  - Choose the feature with the highest/lowest score
  - Re-score other features
  - Repeat
- If you have lots of features (like in text)
  - Just select top K scored features

Unsupervised feature selection

- Differs from feature selection in two ways:
  - Instead of choosing subset of features,
  - Create new features (dimensions) defined as functions over all features
  - Don’t consider class labels, just the data points

Unsupervised feature selection

- Idea:
  - Given data points in d-dimensional space,
  - Project into lower dimensional space while preserving as much information as possible
    - E.g., find best planar approximation to 3D data
    - E.g., find best planar approximation to 104D data
  - In particular, choose projection that minimizes the squared error in reconstructing original data

PCA Algorithm

- PCA algorithm:
  - X ← Create N x d data matrix, with one row vector x_n per data point
  - X subtract mean *x* from each row vector x_n in X
  - Σ ← covariance matrix of X

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}\big[(X_i - \mu_i)(X_j - \mu_j)\big]$$

  - Find eigenvectors and eigenvalues of Σ

$$Av = \lambda v$$

  - PC's ← the M eigenvectors with largest eigenvalues

PCA Algorithm in Matlab

```
% generate data
Data = mvnrnd([5, 5],[1 1.5; 1.5 3], 100);
figure(1); plot(Data(:,1), Data(:,2), '+');
%center the data
for i = 1:size(Data,1)
  Data(i, :) = Data(i, :) - mean(Data);
end

DataCov = cov(Data); %covariance matrix
[PC, variances, explained] = pcacov(DataCov); %eigen

% plot principal components
figure(2); clf; hold on;
plot(Data(:,1), Data(:,2), '+b');
plot(PC(1,1)*[-5 5], PC(2,1)*[-5 5], '-r')
plot(PC(1,2)*[-5 5], PC(2,2)*[-5 5], '-b'); hold off

% project down to 1 dimension
PcaPos = Data * PC(:, 1);
```

2d Data

Principal Components

- Gives best axis to project
- Minimum RMS error
- Principal vectors are orthogonal

How many components?

- Check the distribution of eigen-values
- Take enough eigen-vectors to cover 80-90% of the variance

Sensor networks

Sensors in Intel Berkeley Lab

Pairwise link quality vs. distance

PCA in action

- Given a 54x54 matrix of pairwise link qualities
- Do PCA
- Project down to 2 principal dimensions
- PCA discovered the map of the lab

Eigenfaces

- See demo…

Problems and limitations

- What if very large dimensional data?
  - e.g., Images ($d \geq 10^4$)
- Problem:
  - Covariance matrix Σ is size (d^2)
  - $d=10_4 \rightarrow |\Sigma| = 10^8$
- Singular Value Decomposition (SVD)!
  - efficient algorithms available (Matlab)
  - some implementations find just top N eigenvectors

SVD: Singular Value Decomposition

- A method for transforming correlated variables into a set of uncorrelated ones that better expose the various relationships among the original data items
- SVD identifies and orders the dimensions along which data points exhibit the most variation
- Once we identified those dimensions, it is possible to find the best approximation of the original data using fewer dimensions
- SVD can be seen as a method for data reduction

Singular Value Decomposition

- Problem:
  - #1: Find concepts in text
  - #2: Reduce dimensionality

term / document	data	information	retrieval	brain	lung
CS-TR1	1	1	1	0	0
CS-TR2	2	2	2	0	0
CS-TR3	1	1	1	0	0
CS-TR4	5	5	5	0	0
MED-TR1	0	0	0	2	2
MED-TR2	0	0	0	3	3
MED-TR3	0	0	0	1	1

SVD - Definition

$$\mathbf{A}_{[n \times m]} = \mathbf{U}_{[n \times r]} \, \Lambda_{[r \times r]} \, (\mathbf{V}_{[m \times r]})^T$$

- **A**: *n x m* matrix (e.g., n documents, m terms)
- **U**: *n x r* matrix (n documents, r concepts)
- Λ: *r x r* diagonal matrix (strength of each 'concept') (r: rank of the matrix)
- **V**: m x r matrix (m terms, r concepts)

SVD - Properties

THEOREM [Press+92]: always possible to decompose matrix **A** into $\mathbf{A} = \mathbf{U}\,\Lambda\,\mathbf{V}^T$, where

- **U**, Λ, **V**: unique (*)
- **U**, **V**: column orthonormal (ie., columns are unit vectors, orthogonal to each other)
  - $\mathbf{U}^T\mathbf{U} = \mathbf{I}$; $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ (**I:** identity matrix)
- Λ**:** singular value are positive, and sorted in decreasing order

SVD - Properties

'spectral decomposition' of the matrix:

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} | & | \\ u_1 & u_2 \\ | & | \end{bmatrix} \times \begin{bmatrix} \lambda_1 & \varnothing \\ \varnothing & \lambda_2 \end{bmatrix} \times \begin{bmatrix} \text{———} & v_1 & \text{———} \\ \text{———} & v_2 & \text{———} \end{bmatrix}$$

SVD - Interpretation

‘documents’, ‘terms’ and ‘concepts’:

- **U**: document-to-concept similarity matrix
- **V**: term-to-concept similarity matrix
- Λ: its diagonal elements: ‘strength’ of each concept

Projection:

- best axis to project on: (‘best’ = min sum of squares of projection errors)

SVD - Example

- **A** = **U** Λ **V**^{T} - example:

$$
\begin{array}{c}
\\ \uparrow \\ \text{CS} \\ \downarrow \\ \uparrow \\ \text{MD} \\ \downarrow
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
0.18 & 0 \\
0.36 & 0 \\
0.18 & 0 \\
0.90 & 0 \\
0 & 0.53 \\
0 & 0.80 \\
0 & 0.27
\end{bmatrix}
\times
\begin{bmatrix}
9.64 & 0 \\
0 & 5.29
\end{bmatrix}
\times
\begin{bmatrix}
0.58 & 0.58 & 0.58 & 0 & 0 \\
0 & 0 & 0 & 0.71 & 0.71
\end{bmatrix}
$$

SVD - Example

- **A = U Λ V^T** - example:

doc-to-concept similarity matrix

CS-concept

MD-concept

$$
\begin{array}{c}
\\ \uparrow \\ \text{CS} \\ \downarrow \\ \uparrow \\ \text{MD} \\ \downarrow
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
0.18 & 0 \\
0.36 & 0 \\
0.18 & 0 \\
0.90 & 0 \\
0 & 0.53 \\
0 & 0.80 \\
0 & 0.27
\end{bmatrix}
\times
\begin{bmatrix}
9.64 & 0 \\
0 & 5.29
\end{bmatrix}
\times
\begin{bmatrix}
0.58 & 0.58 & 0.58 & 0 & 0 \\
0 & 0 & 0 & 0.71 & 0.71
\end{bmatrix}
$$

SVD - Example

- **A** = **U** Λ **V**^{T} - example:

'strength' of CS-concept

$$
\begin{array}{c|c}
 & \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\
\begin{array}{c} \text{CS} \\ \\ \text{MD} \end{array} &
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}
\end{array}
=
\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}
\times
\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}
\times
\begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}
$$

SVD - Example

- **A = U Λ V^T** - example:

term-to-concept similarity matrix

$$
\begin{array}{c}
\\ \uparrow \\ \text{CS} \\ \downarrow \\ \uparrow \\ \text{MD} \\ \downarrow
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
0.18 & 0 \\
0.36 & 0 \\
0.18 & 0 \\
0.90 & 0 \\
0 & 0.53 \\
0 & 0.80 \\
0 & 0.27
\end{bmatrix}
\times
\begin{bmatrix}
9.64 & 0 \\
0 & 5.29
\end{bmatrix}
\times
\begin{bmatrix}
0.58 & 0.58 & 0.58 & 0 & 0 \\
0 & 0 & 0 & 0.71 & 0.71
\end{bmatrix}
$$

CS-concept

LSI (latent semantic indexing)

Q1: How to do queries with LSI?

A: map query vectors into ‘concept space’ – how?

$$
\begin{array}{c}
\\ \uparrow \\ \text{CS} \\ \downarrow \\ \uparrow \\ \text{MD} \\ \downarrow
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
0.18 & 0 \\
0.36 & 0 \\
0.18 & 0 \\
0.90 & 0 \\
0 & 0.53 \\
0 & 0.80 \\
0 & 0.27
\end{bmatrix}
\times
\begin{bmatrix}
9.64 & 0 \\
0 & 5.29
\end{bmatrix}
\times
\begin{bmatrix}
0.58 & 0.58 & 0.58 & 0 & 0 \\
0 & 0 & 0 & 0.71 & 0.71
\end{bmatrix}
$$

LSI (latent semantic indexing)

Q: How to do queries with LSI?

A: map query vectors into ‘concept space’ – how?

$$q = \begin{array}{c} \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

A: inner product
(cosine similarity)
with each ‘concept’ vector v_i

LSI (latent semantic indexing)

compactly, we have:

$q_{concept} = q\,\mathbf{V}$

e.g.:

$$q = \begin{array}{c} \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \underbrace{\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}}_{\text{term-to-concept similarities}} = \begin{array}{c} \text{CS-concept} \\ \downarrow \\ \begin{bmatrix} 0.58 & 0 \end{bmatrix} \end{array}$$

Little example

How would the document ('information', 'retrieval') be handled by LSI?

$d_{concept} = d\,\mathbf{V}$

Eg:

$$d = \begin{array}{c} \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix} \end{array} \begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix} = \begin{array}{c} \text{CS-concept} \\ \downarrow \\ \begin{bmatrix} 1.16 & 0 \end{bmatrix} \end{array}$$

term-to-concept similarities

Little example

Observation: document ('information', 'retrieval') will be retrieved by query ('data'), although it does not contain 'data'!!

CS-concept

	data	inf.	retrieval	brain	lung		
d=	0	1	1	0	0	→	1.16 0
q=	1	0	0	0	0	→	0.58 0

Multi-lingual IR

- Solution: ~ LSI

$$
\begin{array}{c}
\\ \uparrow \\ \text{CS} \\ \downarrow \\ \uparrow \\ \text{MD} \\ \downarrow
\end{array}
\begin{array}{c}
\begin{array}{ccccc|ccccc}
\text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} & & & \text{datos} & \text{informacion} & \\
\downarrow & \downarrow & & & & \downarrow & \downarrow & & & \\
1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 & 1 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 & 5 & 5 & 4 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 & 0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 & 0 & 0 & 0 & 2 & 3 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1
\end{array}
\end{array}
$$

- Concatenate documents
- Do SVD on them
- Now when a new document comes project it into concept space
- Measure similarity in concept spalce

ICA

"Independent component analysis (ICA) is a method for finding underlying factors or components from multivariate (multi-dimensional) statistical data. What distinguishes ICA from other methods is that it looks for components that are both *statistically independent*, and *nonGaussian*."

A.Hyvarinen, A.Karhunen, E.Oja

'Independent Component Analysis'

PCA finds the directions of maximum variance

ICA finds the directions of maximum independence

ICA estimation principles

- **Principle 1**: "Nonlinear decorrelation. Find the matrix $\boldsymbol{W}$ so that for any $i \neq j$, the components y_i and y_j are uncorrelated, *and* the transformed components $g(y_i)$ and $h(y_j)$ are uncorrelated, where g and h are some suitable nonlinear functions."
- **Principle 2**: "Maximum nongaussianity". Find the local maxima of nongaussianity of a linear combination $\boldsymbol{y}=\boldsymbol{W}\boldsymbol{x}$ under the constraint that the variance of $\boldsymbol{x}$ is constant.
- Each local maximum gives one independent component.

ICA mathematical approach

"Given a set of observations of random variables $x_1(t)$, $x_2(t) \ldots x_n(t)$, where t is the time or sample index, assume that they are generated as a linear mixture of independent components: $\boldsymbol{y}=\boldsymbol{W}\boldsymbol{x}$, where $\boldsymbol{W}$ is some unknown matrix.

Independent component analysis now consists of estimating both the matrix $\boldsymbol{W}$ and the $y_i(t)$, when we only observe the $x_i(t)$."

Principle: Maximize Information

- **Q:** How to extract **maximum information** from multiple visual channels**?**

- **A**: ICA does this -- it **maximizes** joint entropy **& minimizes** mutual information between output channels (Bell & Sejnowski, 1995).
- ICA produces brain-like visual filters for natural images.

Set of 144 ICA filters

Example: Audio decomposition

Perform ICA

Play Mixtures

Terry Scott

Te-Won Tzyy-Ping

Play Components

ICA Applications

- Sound source separation
- Image processing
- Sonar target identification
- Underwater communications
- Wireless communications
- Brain wave analysis (EEG)
- Brain imaging (fMRI)

Recordings in real environments

Separation of Music & Speech

Experiment-Setup:

- *office room (5m x 4m)*
- *two distant talking mics*
- *16kHz sampling rate*

Learning Image Features

y Picture-Perfect Slice of th

ing on Route 4 over a dark and winding mountain pass, the visitor suddenly emerges into a "Lost Horizon" world of hot springs, trout streams and meadows of wildflowers, where cattle and the state's largest elk herd graze side by side.

But the same sense of wide-open Western independence evoked by the vistas has prevented the sale of the land for years. And the deal that is being negotiated for the ranch, which has been owned by one family for almost 40 years, is as much about Western attitudes toward public land as it is about money.

The Administration has long supported the purchase of the ranch, which has been called "the hole in the doughnut" because it is an island surrounded by the Santa Fe National Forest. Last February on a visit to New Mexico, Preside had Air Force One m fly over the ranch for its dominant feature wide crater of the do

Republicans, notin one-third of New Me owned by the Feders have long opposed chases. But in Augu sentiment began shi cally.

Under legislation d tor Pete V. Domenici New Mexico, the Ba separate unit of the system, owned by th Forest Service, but trust, comprised of pointed by the Presid

Continued on P

Learning Image Features

Automatic Image Segmentation

JUVENILE KOMODOS spend much of their time in trees, a behavior that keeps them from being eaten by larger, less agile members of their species.

suddenly launches the coup de grâce, a deadly attack. The deer quickly bleeds to death, and the Komodo begins to feed.

The muscles of the Komodo's jaws and throat allow it to swallow huge chunks of meat with astonishing rapidity: Auffenberg once observed a female who weighed no more than 50 kilograms consume a 31-kilogram boar in 17 minutes. Several movable joints, such as the intramandibular hinge that opens the lower jaw unusually wide, help in the bolting. The stomach expands easily, enabling an adult to consume up to 80 percent of its own body weight in a single meal, which most likely explains some exaggerated claims for immense weights in captured individuals.

Large mammalian carnivores, such as lions, tend to leave 25 to 30 percent of their kill unconsumed, declining the intestines, hide, skeleton and hooves. Komodos eat much more efficiently, forsaking only about 12 percent of the prey. They eat bones, hooves and swaths of hide. They also eat intestines, but only after swinging them vigorously to scatter their contents. This behavior removes feces from the meal. Because large Komodos cannibalize young ones, the latter often roll in fecal material, thereby assuming a scent that their bigger brethren are programmed to avoid consuming.

More Komodos, attracted by the aromas, arrive and join in the feeding. Although males tend to grow larger and bulkier than females, no obvious morphological differences mark the sexes. One subtle clue does exist: a slight difference in the arrangement of scales just in front of the cloaca, the cavity housing the genitalia in both sexes. Sexing Komodos remains a challenge to researchers; the dragons themselves appear to have little trouble figuring out who is who. With a group assembled around the carrion, the opportunity for courtship arrives.

Most mating occurs between May and August. Dominant males can become embroiled in ritual combat in their quest for females. Using their tails for support, they wrestle in upright postures, grabbing each other with their forelegs as they attempt to throw the opponent to the ground. Blood is usually drawn, and the loser either runs or remains prone and motionless.

The victorious wrestler initiates courtship by flicking his tongue on a female's snout and then over her body. The temple and the fold between the torso and the rear leg are favorite spots. Stimula-

Barcode Classification

Matrix

Linear

Postal

Learned ICA Output Filters

Matrix

Postal

Linear

Image De-noising

Filling in missing data

ICA applied to Brainwaves

An EEG recording consists of activity arising from many brain and extra-brain processes

Time Course of ICA Components
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
0
1
2
3
4
5
Time(sec)
Eye movement
Muscle activity

Original EEG
Corrected EEG
Fp1
Fp2
F3
F4
C3
C4
A2
P3
P4
O1
O2
F7
F8
T3
T4
T5
T6
Fz
Cz
Pz
EOG1
EOG2
0
1
2
3
4
5
Time(sec)

Single trials
40
20
0
−20
−40
Amplitude (μV)

Single-trial ERPs at Cz
Trial Number
0
50
100
150
200
250
300
350
21
μV
-7
Median RT
-100
100
300
500
700
900
Time (msec)

Artifacts
Stimulus–locked
Response–locked
Not phase–locked
Alpha Phase–resetting
Mu Blocking
RT–Sorted Trial Index
Time (msec)
0
200
400
600
800
–500
500
1000
1500
CoRms
11 Hz