Reducing Data Dimension

Recommended readings (available on class website):

• “A Tutorial on PCA,’ J. Schlens
• Wall et al., 2003
Outline

• Unsupervised dimension reduction using all features
  – Principle Components Analysis
  – Singular Value Decomposition
  – Independent components analysis
  – Canonical correlation analysis

• Supervised dimension reduction
  – Fisher Linear Discriminant
  – Hidden layers of Neural Networks
Unsupervised Dimensionality Reduction
Principle Components Analysis

• 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 $10^4$ D data
  – In particular, choose projection that minimizes the squared error in reconstructing original data
PCA: Find Projections to Minimize Reconstruction Error

Assume data is set of d-dimensional vectors, where nth vector is 
\[ x^n = \sum_{i=1}^{d} u_i n \]

We can represent these in terms of any d orthogonal vectors \( u_1 \ldots u_d \)
\[ x^n = \sum_{i=1}^{d} z^n_i u_i; \quad u_i^T u_j = \delta_{ij} \]

PCA: given M<d. Find \( \langle u_1 \ldots u_M \rangle \)

that minimizes 
\[ E_M = \sum_{n=1}^{N} ||x^n - \hat{x}^n||^2 \]

where \( \hat{x}^n = \bar{x} + \sum_{i=1}^{M} z^n_i u_i \)

\[ \bar{x} = \frac{1}{N} \sum_{n=1}^{N} x^n \]
PCA: given $M<d$. Find $\langle u_1 \ldots u_M \rangle$

that minimizes $E_M \equiv \sum_{n=1}^{N} ||x^n - \hat{x}^n||^2$

where $\hat{x}^n = \bar{x} + \sum_{i=1}^{M} z_i^n u_i$

Note we get zero error if $M=d$, so all error is due to missing components.

Therefore,

$E_M = \sum_{i=M+1}^{d} \sum_{n=1}^{N} [u_i^T (x^n - \bar{x})]^2$

$= \sum_{i=M+1}^{d} u_i^T \Sigma u_i$

Covariance matrix: $\Sigma = \sum_{n} (x^n - \bar{x})(x^n - \bar{x})^T$

This minimized when $u_i$ is eigenvector of $\Sigma$, the covariance matrix of $X$. i.e., minimized when:

$\Sigma u_i = \lambda_i u_i$
Minimize $E_M = \sum_{i=M+1}^{d} u_i^T \Sigma u_i$

$\implies \Sigma u_i = \lambda_i u_i$

Eigenvector of $\Sigma$

Eigenvalue (scalar)

$\implies E_M = \sum_{i=M+1}^{d} \lambda_i$

PCA algorithm 1:

1. $X \leftarrow$ Create $N \times d$ data matrix, with one row vector $x^n$ per data point
2. $X \leftarrow$ subtract mean $\bar{x}$ from each row vector $x^n$ in $X$
3. $\Sigma \leftarrow$ covariance matrix of $X$
4. Find eigenvectors and eigenvalues of $\Sigma$
5. $\text{PC’s} \leftarrow$ the $M$ eigenvectors with largest eigenvalues
PCA Example

\[ \hat{x}^n = \bar{x} + \sum_{i=1}^{M} z_i^n u_i \]
PCA Example

\[ \hat{x}^n = \bar{x} + \sum_{i=1}^{M} z_i^n u_i \]

Reconstructed data using only first eigenvector (M=1)
Very Nice When Initial Dimension Not Too Big

What if very large dimensional data?
• e.g., Images (d \approx 10^4)

Problem:
• Covariance matrix $\Sigma$ is size $(d \times d)$
• $d=10^4 \rightarrow |\Sigma| = 10^8$

Singular Value Decomposition (SVD) to the rescue!
• pretty efficient algs available, including Matlab SVD
• some implementations find just top N eigenvectors
SVD

Data $X$, one row per data point

$US$ gives coordinates of rows of $X$ in the space of principle components

$S$ is diagonal, $S_k > S_{k+1}$, $S_k^2$ is kth largest eigenvalue

Rows of $V^T$ are unit length eigenvectors of $X^TX$

If cols of $X$ have zero mean, then $X^TX = c \Sigma$ and eigenvects are the Principle Components

[from Wall et al., 2003]
Singular Value Decomposition

To generate principle components:

• Subtract mean \( \bar{x} = \frac{1}{N} \sum_{n=1}^{N} x^n \) from each data point, to create zero-centered data
• Create matrix \( X \) with one row vector per (zero centered) data point
• Solve SVD: \( X = USV^T \)
• Output Principle components: columns of \( V (= \text{rows of } V^T) \)
  – Eigenvectors in \( V \) are sorted from largest to smallest eigenvalues
  – \( S \) is diagonal, with \( s_k^2 \) giving eigenvalue for \( k \)th eigenvector
Singular Value Decomposition

To project a point (column vector $x$) into PC coordinates:

$$V^T x$$

If $x_i$ is $i$th row of data matrix $X$, then

- $(i$th row of $US) = V^T x_i^T$
- $(US)^T = V^T X^T$

To project a column vector $x$ to M dim Principle Components subspace, take just the first M coordinates of $V^T x$
PCA Example

faces

Eigenfaces

1st
2nd

Thanks to Christopher DeCoro
see http://www.cs.princeton.edu/~cdecoro/eigenfaces/
Reconstructing a face from the first $N$ components (eigenfaces)

Adding 1 additional PCA component at each step

Adding 8 additional PCA components at each step

In this next image, we show a similar picture, but with each additional face representing an additional 8 principle components. You can see that it takes a rather large number of images before the picture looks totally correct.
Supervised Dimensionality Reduction
1. Hidden Layers in Neural Networks

When \# hidden units < \# inputs, hidden layer also performs dimensionality reduction.

Each synthesized dimension (each hidden unit) is logistic function of inputs

\[
h_k(x) = \frac{1}{1 + exp(w_0 + \sum_{i=1}^{N} w_i x_i)}
\]

Hidden units defined by gradient descent to (locally) minimize squared output classification/regression error

\[
E = \sum_{n=1}^{N} \sum_{k} (\hat{y}_k(x^n) - y_k(x^n))^2
\]

Also allow networks with multiple hidden layers
→ highly nonlinear components (in contrast with linear subspace of Fisher LD, PCA)
Learning Hidden Layer Representations

Training neural network to minimize reconstruction error

A target function:

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000000 → 10000000</td>
<td></td>
</tr>
<tr>
<td>01000000 → 01000000</td>
<td></td>
</tr>
<tr>
<td>00100000 → 00100000</td>
<td></td>
</tr>
<tr>
<td>00010000 → 00010000</td>
<td></td>
</tr>
<tr>
<td>00001000 → 00001000</td>
<td></td>
</tr>
<tr>
<td>00000100 → 00000100</td>
<td></td>
</tr>
<tr>
<td>00000010 → 00000010</td>
<td></td>
</tr>
<tr>
<td>00000001 → 00000001</td>
<td></td>
</tr>
</tbody>
</table>

Can this be learned??
Learning Hidden Layer Representations

A network:

Learned hidden layer representation:

<table>
<thead>
<tr>
<th>Input</th>
<th>Hidden Values</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>100000000</td>
<td>.89 .04 .08</td>
<td>100000000</td>
</tr>
<tr>
<td>010000000</td>
<td>.01 .11 .88</td>
<td>010000000</td>
</tr>
<tr>
<td>001000000</td>
<td>.01 .97 .27</td>
<td>001000000</td>
</tr>
<tr>
<td>000100000</td>
<td>.99 .97 .71</td>
<td>000100000</td>
</tr>
<tr>
<td>000010000</td>
<td>.03 .05 .02</td>
<td>000010000</td>
</tr>
<tr>
<td>000001000</td>
<td>.22 .99 .99</td>
<td>000001000</td>
</tr>
<tr>
<td>000000100</td>
<td>.80 .01 .98</td>
<td>000000100</td>
</tr>
<tr>
<td>000000010</td>
<td>.60 .94 .01</td>
<td>000000010</td>
</tr>
</tbody>
</table>
Neural Nets for Face Recognition

Typical input images

90% accurate learning head pose, and recognizing 1-of-20 faces
Learned Hidden Unit Weights

left strt rght up

Learned Weights

30x32 inputs

Typical input images

http://www.cs.cmu.edu/~tom/faces.html
2. Fisher Linear Discriminant

- A method for projecting data into lower dimension to hopefully improve classification

- We’ll consider 2-class case

Project data onto vector that connects class means?
Summary: Fisher Linear Discriminant

- Choose n-1 dimension projection for n-class classification problem
- Use within-class covariances to determine the projection
- Minimizes a different sum of squared error function (the projected within-class variances)
Fisher Linear Discriminant

Project data onto one dimension, to help classification

\[ y = w^T x \]

Define class means:

\[ m_i = \frac{1}{N_i} \sum_{n \in C_i} x^n \]

Could choose w according to:

\[ \arg \max_w w^T (m_2 - m_1) \]

Instead, Fisher Linear Discriminant chooses:

\[ \arg \max_w \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \]

\[ m_i \equiv w^T m_i \quad s_i^2 \equiv \sum_{n \in C_i} (x^n - m_i)^2 \]
Fisher Linear Discriminant

Project data onto one dimension, to help classification

\[ y = w^T x \]

Fisher Linear Discriminant: \( \arg \max_w \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \)

is solved by: \( w \propto S_W^{-1}(m_2 - m_1) \)

Where \( S_W \) is sum of within-class projected covariances:

\[
S_W = \sum_{n \in C_1} (x^n - m_1)(x^n - m_1)^T + \sum_{n \in C_2} (x^n - m_2)(x^n - m_2)^T
\]
Fisher Linear Discriminant

Fisher Linear Discriminant: \[ \arg \max_w \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \]

Is equivalent to minimizing sum of squared error if we assume target values are not +1 and -1, but instead \( \frac{N}{N_1} \) and \( -\frac{N}{N_2} \)

Where \( N \) is total number of examples, \( N_i \) is number in class i

Also generalized to K classes (and projects data to K-1 dimensions)
Summary: Fisher Linear Discriminant

- Choose n-1 dimension projection for n-class classification problem
- Use within-class covariances to determine the projection
- Minimizes a different sum of squared error function (the projected within-class variances)
Even More Approaches
Dimensionality reduction for multiple datasets

• Given data sets A and B, find linear projections of each into a common lower dimensional space
  – Generalized SVD: to minimize sq reconstruction errors of both
  – Canonical correlation analysis: to maximize correlation of A and B in the abstract space
Canonical Correlation Analysis
Component 1

Top + verbs: bite chop cut tip ring grip hold bend dry remove

Top - verbs: build travel plan unite repair see sleep open walk live

[Rustandi, 2009]
Top + verbs: open lift stand hold break work tip drop pull bare
Top - verbs: plant chop paste build dry treat ring eat love rescue
Bags of Words, or Bags of Topics?
Example topics
induced from a large collection of text

[DISEASE
BACTERIA
DISEASES
GERMS
FEVER
CAUSE
CAUSED
SPREAD
VIRUSES
INFECTION
VIRUS
INFECTION
PERSON
INFECTIOUS
COMMON
CAUSING
SMALLPOX
BODY
INFECTIONS
CERTAIN
MIND
WORLD
WATER
FISH
SEA
SWIM
STORY
STORIES
TELL
CHARACTER
CHARACTERS
AUTHOR
READ
TOLD
SETTING
TALES
TELLING
SHORT
FICTION
NOVEL
FIELD
MAGNETIC
MAGNET
WIRE
NEEDLE
COIL
POLES
IRON
COMPASS
LINES
CORE
ELECTRIC
DIRECTION
FORCE
MAGNETS
BE
MAGNETISM
POLE
INDUCED
FIELD
MAGNETIC
MAGNET
WIRE
NEEDLE
COIL
POLES
IRON
COMPASS
LINES
CORE
ELECTRIC
DIRECTION
FORCE
MAGNETS
BE
MAGNETISM
POLE
INDUCED
SCIENCE
STUDY
SCIENTISTS
SCIENTIFIC
KNOWLEDGE
WORK
RESEARCH
CHEMISTRY
TECHNOLOGY
MANY
MATHEMATICS
BIOLOGY
FIELD
PHYSICS
LABORATORY
STUDIES
WORLD
SCIENTIST
STUDYING
SCiences
BALL
GAME
TEAM
FOOTBALL
BASEBALL
PLAYERS
PLAY
FIELD
PLAYER
BASKETBALL
COACH
PLAYED
PLAYING
HIT
TENNIS
TEAMS
GAMES
SPORTS
BAT
TERRY
JOB
WORK
JOBS
CAREER
EXPERIENCE
EMPLOYMENT
OPPORTUNITIES
WORKING
TRAINING
SKILLS
CAREERS
POSITIONS
FIND
POSITION
FIELD
OCCUPATIONS
REQUIRE
OPPORTUNITY
EARN
ABLE

[Tennenbaum et al]
Clustering words into topics with Hierarchical Topic Models (unknown number of clusters) [Blei, Ng, Jordan 2003]

Probabilistic model for generating document D:

1. Pick a $\theta \sim P(\theta|\alpha)$ to define $P(z|\theta)$
2. For each of the $N_d$ words $w$
   - Pick topic $z$ from $P(z | \theta)$
   - Pick word $w$ from $P(w | z, \phi)$

Training this model defines topics (i.e., $\phi$ which defines $P(W|Z)$)
Latent Dirichlet Allocation Model

Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

where $p(z_n | \theta)$ is simply $\theta_i$ for the unique $i$ such that $z_n^i = 1$. Integrating over $\theta$ and summing over $z$, we obtain the marginal distribution of a document:

$$p(w | \alpha, \beta) = \int p(\theta | \alpha) \left( \prod_{n=1}^{N} \sum_{z_n} p(z_n | \theta)p(w_n | z_n, \beta) \right) d\theta. \quad (3)$$
Example topics induced from a large collection of text

Tennenbaum et al
Example topics induced from a large collection of text

Significance:

• Learned topics reveal implicit semantic categories of words within the documents

• In many cases, we can represent documents with $10^2$ topics instead of $10^5$ words

• Especially important for short documents (e.g., emails). Topics overlap when words don’t!

[tennenbaum et al]
Analyzing topic distributions in email
Author-Recipient-Topic model for Email

Latent Dirichlet Allocation (LDA)
[Blei, Ng, Jordan, 2003]

Author-Recipient Topic (ART)
[McCallum, Corrada, Wang, 2005]
Date: Wed, 11 Apr 2001 06:56:00 -0700 (PDT)
From: debra.perlingiere@enron.com
To: steve.hooser@enron.com
Subject: Enron/TransAlta Contract dated Jan 1, 2001

Please see below. Katalin Kiss of TransAlta has requested an electronic copy of our final draft? Are you OK with this? If so, the only version I have is the original draft without revisions.

DP

Debra Perlingiere
Enron North America Corp.
Legal Department
1400 Smith Street, EB 3885
Houston, Texas 77002
dperlin@enron.com
Topics, and prominent sender/receivers discovered by ART

[McCallum et al, 2005]

<table>
<thead>
<tr>
<th>Topic 17</th>
<th>Topic 27</th>
<th>Topic 45</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>“Document Review”</strong></td>
<td><strong>“Time Scheduling”</strong></td>
<td><strong>“Sports Pool”</strong></td>
</tr>
<tr>
<td>attached</td>
<td>day</td>
<td>game</td>
</tr>
<tr>
<td>agreement</td>
<td>friday</td>
<td>draft</td>
</tr>
<tr>
<td>review</td>
<td>morning</td>
<td>week</td>
</tr>
<tr>
<td>questions</td>
<td>monday</td>
<td>team</td>
</tr>
<tr>
<td>draft</td>
<td>office</td>
<td>eric</td>
</tr>
<tr>
<td>letter</td>
<td>wednesday</td>
<td>make</td>
</tr>
<tr>
<td>comments</td>
<td>tuesday</td>
<td>free</td>
</tr>
<tr>
<td>copy</td>
<td>time</td>
<td>year</td>
</tr>
<tr>
<td>revised</td>
<td>good</td>
<td>pick</td>
</tr>
<tr>
<td>document</td>
<td>thursday</td>
<td>phillip</td>
</tr>
</tbody>
</table>

Top words within topic:

- attached
- agreement
- review
- questions
- draft
- letter
- comments
- copy
- revised
- document

Top author-recipients exhibiting this topic:

- G. Nemec
- B. Tycholiz
- J. Dasovich
- R. Shapiro
- E. Bass
- M. Lenhart
- M. Motley
- M. Grigsby

- G. Nemec
- M. Whitt
- J. Dasovich
- J. Steffes
- E. Bass
- P. Love

- B. Tycholiz
- G. Nemec
- C. Clair
- M. Taylor
- M. Motley
- M. Grigsby
Topics, and prominent sender/receivers discovered by ART

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>operations 0.0321</td>
<td>market 0.0567</td>
<td>state 0.0404</td>
<td>blackberry 0.0726</td>
</tr>
<tr>
<td>team 0.0234</td>
<td>power 0.0563</td>
<td>california 0.0367</td>
<td>net 0.0557</td>
</tr>
<tr>
<td>office 0.0173</td>
<td>price 0.0280</td>
<td>power 0.0337</td>
<td>www 0.0409</td>
</tr>
<tr>
<td>list 0.0144</td>
<td>system 0.0206</td>
<td>energy 0.0239</td>
<td>website 0.0375</td>
</tr>
<tr>
<td>bob 0.0129</td>
<td>prices 0.0182</td>
<td>electricity 0.0203</td>
<td>report 0.0373</td>
</tr>
<tr>
<td>open 0.0126</td>
<td>high 0.0124</td>
<td>davis 0.0183</td>
<td>wireless 0.0364</td>
</tr>
<tr>
<td>meeting 0.0107</td>
<td>based 0.0120</td>
<td>utilities 0.0158</td>
<td>handheld 0.0362</td>
</tr>
<tr>
<td>gas 0.0107</td>
<td>buy 0.0117</td>
<td>commission 0.0136</td>
<td>stan 0.0282</td>
</tr>
<tr>
<td>business 0.0106</td>
<td>customers 0.0110</td>
<td>governor 0.0132</td>
<td>fyi 0.0271</td>
</tr>
<tr>
<td>houston 0.0099</td>
<td>costs 0.0106</td>
<td>prices 0.0089</td>
<td>named 0.0260</td>
</tr>
</tbody>
</table>

S.Beck 0.2158 L.Kitchen 0.1231 J.Dasovich 0.1231 | J.Dasovich 0.3338 R.Haylett 0.1432 T.Geaccone
S.Beck 0.0826 J.Lavorato 0.1133 | J.Dasovich 0.2440 T.Geaccone 0.0737 R.Haylett
S.Beck 0.0530 S.White 0.0218 | M.Taylor 0.0218 J.Dasovich 0.1394 R.Haylett 0.0420 D.Fossum

Beck = “Chief Operations Officer”
Dasovitch = “Government Relations Executive”
Shapiro = “Vice Presidency of Regulatory Affairs”
Steffes = “Vice President of Government Affairs”
Discovering Role Similarity

**Traditional SNA**

connection strength \((A,B) = \)

Similarity in recipients they sent email to

**ART**

Similarity in authored topics, conditioned on recipient
Discovering Role Similarity

Tracy Geaconne ⇔ Dan McCarty

- **Traditional SNA**
  - Similar: (send email to same individuals)
  - Different: (discuss different topics)

- **ART**

  Geaconne = “Secretary”
  McCarty = “Vice President”
Discovering Role Similarity
Lynn Blair ↔ Kimberly Watson

Traditional SNA

ART

Different
(send to different individuals)

Similar
(discuss same topics)

Blair = “Gas pipeline logistics”
Watson = “Pipeline facilities planning”
Independent Components Analysis

- PCA seeks orthogonal directions $<Y_1 \ldots Y_M>$ in feature space $X$ that minimize reconstruction error.

- ICA seeks directions $<Y_1 \ldots Y_M>$ that are most statistically independent. I.e., that minimize $I(Y)$, the mutual information between the $Y_j$:

$$I(Y) = \left[ \sum_{j=1}^{J} H(Y_j) \right] - H(Y)$$
What you should know

• Unsupervised dimension reduction using all features
  – Principle Components Analysis
    • Minimize reconstruction error
  – Singular Value Decomposition
    • Efficient PCA
  – Independent components analysis
  – Canonical correlation analysis
  – Probabilistic models with latent variables

• Supervised dimension reduction
  – Fisher Linear Discriminant
    • Project to n-1 dimensions to discriminate n classes
  – Hidden layers of Neural Networks
    • Most flexible, local minima issues

• LOTS of ways of combining discovery of latent features with classification tasks
Further Readings