Lecture 10. Big Data Classification by Principal Component Analysis

Muhammad A. Alam
alam@purdue.edu
This material is copyrighted by M. Alam under the following Creative Commons license:

Conditions for using these materials is described at

http://creativecommons.org/licenses/by-nc-sa/2.5/
Course Outline

\[ \bar{y} = f(\bar{x}) \quad \bar{x} = x_1, x_2, \ldots x_n \quad \bar{y} = y_1, y_2, \ldots y_m \]

Lecture 1: Introduction
Lecture 2: Collecting and plotting \( x_1, x_2, \ldots x_n \)
Lecture 3: Physical and empirical \( f, F, df/dx, \ldots \)
Lecture 4: Model selection between \( f_1, f_2, \ldots \)
Lecture 5: Model Selection: Cross-validation and Bootstrapping method
Lecture 6: Scaling theory with known \( f, f(\bar{x}) = f(\bar{X}) \)
Lecture 7: Scaling theory with unknown \( f, \bar{x} \rightarrow X \)
Lecture 8: Design of experiments to determine \( \bar{y}_{\text{max}} = f(\bar{x}) \)
Lecture 9: DOE and ANOVA
Lecture 10: Principle component analysis for classifying \{y\}.  
Lecture 11: Machine learning … Statistical approach to learn \( f \)
Lecture 12: Interpretable ML: Physics-based machine learning \( f = f_{\text{physics}} + \Delta f \)
Lecture 13: Interpretable ML: System Equation Modeling
Lecture 14: Conclusions
Big vs. small data

• Big data is obtained as is. One must ask intelligent questions to tease-out the answers embedded within the information. Census and insurance information are examples. Analysis is difficult, but they do represent real world conditions.

• Small data is often hypothesis driven and obtained from carefully designed experiments or survey. Data acquisition is planned and therefore expensive. The analysis is simpler, but may not represent real world conditions.
Small vs. big data
Where do data come from?

- Hundreds of petabyte of data every day.
- Social media sites
- Digital pictures
- Videos
- Purchase transaction
- GPS signals and so on.
- Scientific instrumentation
- Census data
... driven by memory technology

- Cisco estimates: 1.8 ZB by 2016 and 7.2 ZB in 2021.
- If 1 MB is the size of the period at the end of sentence, 1.8 ZB is 460 km^2, eight times the size of Manhattan
- Amazon Web services, Google Cloud, IBM Cloud, Microsoft Azure.

<table>
<thead>
<tr>
<th>Storage Type</th>
<th>Access Time</th>
<th>Capacity</th>
<th>Data Persistence</th>
<th>Read/Write Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solid State Drive</td>
<td>50/1000 ns</td>
<td>2 terabytes</td>
<td>8-10 years</td>
<td>1000</td>
</tr>
<tr>
<td>Hard-Disk Drive</td>
<td>7 millisecond</td>
<td>8 terabytes</td>
<td>3-6 years</td>
<td>Indefinite</td>
</tr>
<tr>
<td>Magnetic Tape</td>
<td></td>
<td>12 terabytes</td>
<td>10-30 years</td>
<td>Indefinite</td>
</tr>
</tbody>
</table>
“Big data” techniques apply to “little data” too

Isaac Newton, known as a physicist, mathematician and astronomer, may have also been the “cat door inventor”! According to an anecdote, Newton foolishly made a large hole for the mother cat and six small holes for her six kittens, not understanding that the kittens could follow their mother through the large hole!
Our goal for the next few lectures ...

Lectures 6-7
\[ \bar{x} = x_1, x_2, \ldots, x_n \]

\[ \bar{y} = f(\bar{x}) \]

Lectures 3-5

Lectures 8-14
How to get a better \( f \)

Lectures 1-2
\[ \bar{y} = y_1, y_2, \ldots, y_m \]

How to fit multiple hypothetical function \( f \) to the same \( y \)
Analysis of big data

Data Processing
- Distance Metrics (2.1)
- Sampling (2.2)
- Dimensionality Reduction (2.3)
- PCA (2.3.1) and SVD (2.3.2)

Model Learning
- Supervised
  - Regression, Classifications
    - kNN (3.1.1), Decision Trees (3.1.2), Rules (3.1.3), Bayesian Classification (3.1.4), Logistic Regression (3.1.5), SVM (3.1.5), and ANN (3.1.7)
- Unsupervised Learning
  - Associated Rules, Matrix Completion, Clustering
  - k-means (4.1.1), Density-based (4.1.2), Message passing (4.1.2), Hierarchical (4.1.2), LDA (4.1.2), Bayesian Non-parametric (4.1.2), LSH (4.1.2)

Testing and Validation
- Evaluating classifiers (3.3)
Outline

1. Introduction

2. Why do we need reduction in data dimension

2. Theory of Principle Component Analysis

3. Applications of Principle Component Analysis

4. Conclusions
Classification problem in big data

Advertisement
Recommendation

Facial Recognition
Voice Recognition
Spam Filtering

Everything is a Recommendation

Over 75% of what people watch comes from our recommendations

Recommendations are driven by Machine Learning
PCA helps classification

<table>
<thead>
<tr>
<th>name</th>
<th>100m</th>
<th>Long Jump</th>
<th>Javeline</th>
<th>2500m</th>
<th>Rank</th>
<th>Points</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ville</td>
<td>11.06</td>
<td>7.08</td>
<td>64.19</td>
<td>291.7</td>
<td>1</td>
<td>8277</td>
<td>Decentor</td>
</tr>
<tr>
<td>Candy</td>
<td>10.76</td>
<td>7.4</td>
<td>60.15</td>
<td>301.5</td>
<td>2</td>
<td>8332</td>
<td>Decentor</td>
</tr>
<tr>
<td>Money</td>
<td>10.80</td>
<td>7.47</td>
<td>59.46</td>
<td>299.42</td>
<td>3</td>
<td>8414</td>
<td>Olympian</td>
</tr>
<tr>
<td>Marois</td>
<td>10.62</td>
<td>7.74</td>
<td>55.89</td>
<td>276.91</td>
<td>5</td>
<td>8343</td>
<td>Olympian</td>
</tr>
<tr>
<td>V.</td>
<td>10.91</td>
<td>7.14</td>
<td>63.49</td>
<td>269.34</td>
<td>6</td>
<td>8287</td>
<td>Olympian</td>
</tr>
<tr>
<td>Kento</td>
<td>10.97</td>
<td>7.19</td>
<td>57.79</td>
<td>264.35</td>
<td>7</td>
<td>8237</td>
<td>Olympian</td>
</tr>
<tr>
<td>Rappeneau</td>
<td>10.95</td>
<td>7.31</td>
<td>53.45</td>
<td>287.43</td>
<td>11</td>
<td>8084</td>
<td>Olympian</td>
</tr>
<tr>
<td>Schonebeek</td>
<td>10.95</td>
<td>7.3</td>
<td>66.89</td>
<td>278.82</td>
<td>12</td>
<td>8077</td>
<td>Olympian</td>
</tr>
<tr>
<td>Revas</td>
<td>11.14</td>
<td>6.99</td>
<td>66.55</td>
<td>267.89</td>
<td>13</td>
<td>8067</td>
<td>Olympian</td>
</tr>
<tr>
<td>Mionnet</td>
<td>11.02</td>
<td>7.1</td>
<td>50.31</td>
<td>266.2</td>
<td>15</td>
<td>8096</td>
<td>Decentor</td>
</tr>
<tr>
<td>Marinier</td>
<td>11.13</td>
<td>7.6</td>
<td>51.97</td>
<td>278.5</td>
<td>19</td>
<td>8030</td>
<td>Decentor</td>
</tr>
<tr>
<td>Noel</td>
<td>10.88</td>
<td>7.03</td>
<td>61.83</td>
<td>276.33</td>
<td>8</td>
<td>8236</td>
<td>Olympian</td>
</tr>
<tr>
<td>Dixas</td>
<td>10.97</td>
<td>7.01</td>
<td>51.61</td>
<td>274.21</td>
<td>19</td>
<td>7936</td>
<td>Olympian</td>
</tr>
</tbody>
</table>

Active individuals
Active variables
Supplementary quantitative variables
Supplementary qualitative variable
Supplementary individuals
PCA Also help in data compression

3D information projected onto a 2D plane

Perspective projection invented by Fillipo Brunelleschi & Masaccio
Outline

1. Why do we need reduction in data dimension
2. Theory of Principle Component Analysis
3. Applications of Principle Component Analysis
4. Conclusions
Principle Component Analysis (PCA)

- Example: Are some students falling behind? Difficult to decide in a multidimensional data

<table>
<thead>
<tr>
<th>Student</th>
<th>Math</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97</td>
<td>94</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>89</td>
<td>76</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>45</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>90</td>
<td>86</td>
</tr>
</tbody>
</table>

2 measurements

<table>
<thead>
<tr>
<th>Student</th>
<th>PC1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>

1 measurement
Basic Concept of PCA

To reduce 2-D data to 1-D data: find a direction onto which to project the data so as to minimize the projection error.

Projection error = $\sum_N a_N^2$
PCA through Singular Value Decomposition

\[ X = U \Sigma V^T \]

- **\( X \):** \( n \times m \) matrix
- **\( U \):** \( n \times n \) matrix
- **\( \Sigma \):** \( n \times m \) diagonal matrix
- **\( V \):** \( m \times m \) matrix

**Direction of the new axis (PC1, PC2 …)**

**Weight of each principal components**

<table>
<thead>
<tr>
<th>Student</th>
<th>Math</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97</td>
<td>94</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>89</td>
<td>76</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>45</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>( N )</td>
<td>90</td>
<td>86</td>
</tr>
</tbody>
</table>
Reduce dimension by Singular Value Decomposition
Example 1: Rotation matrix

\[ R = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \]

\[ R^{90} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \]

\[ R^{180} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \end{pmatrix} \]

\[ R^{270} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \]
SVD rotates the axes optimally

Three points \((1,1), (2,2)\) and \((0,0)\)

\[ X = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 0 & 0 \end{pmatrix} \]

SVD \(\{\{1,1\}, \{2,2\}, \{0,0\}\}\)

Rotate by 45 degrees
The PC is sufficient

\[ V^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \]

\[ U = \begin{pmatrix} \frac{1}{\sqrt{5}} & 0 & -\frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & 0 & \frac{1}{\sqrt{5}} \\ 0 & 1 & 0 \end{pmatrix} \]

\[ \Sigma = \begin{pmatrix} \sqrt{10} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \]
SVD components allows reconstruction

\[
X = \begin{pmatrix}
1 & 1 \\
2 & 2 \\
0 & 0
\end{pmatrix} = u_1 \sigma_1 v_1^T + u_2 \sigma_2 v_2^T
\]

\[
u_1 = \begin{pmatrix}
1/\sqrt{5} \\
2/\sqrt{5} \\
0
\end{pmatrix}
\]

\[
\sigma_1 = \sqrt{10} \quad v_1^T = \begin{pmatrix}
1/\sqrt{2} \\
1/\sqrt{2}
\end{pmatrix}
\]

\[
X = \begin{pmatrix}
1 & 1 \\
2 & 2 \\
0 & 0
\end{pmatrix} \quad X' = u_1 \sigma_1 v_1^T = \begin{pmatrix}
1 & 1 \\
2 & 2 \\
0 & 0
\end{pmatrix}
\]

\[
X' = X \text{ because the projection is exact}
\]
Projection along PCs

\[ XV = U \Sigma V^T V = U \Sigma \]

\[ X = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 0 & 0 \end{pmatrix} \]

\[ V = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \]

\[ U = \begin{pmatrix} 1/\sqrt{5} & 0 & -2/\sqrt{5} \\ 2/\sqrt{5} & 0 & 1/\sqrt{5} \\ 0 & 1 & 0 \end{pmatrix} \]

\[ \Sigma = \begin{pmatrix} \sqrt{10} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \]

\[ U.\Sigma = \begin{pmatrix} \sqrt{2} & 0 \\ 2\sqrt{2} & 0 \\ 0 & 0 \end{pmatrix} \]

Projection along PC1
Example 2: More general result

Three points \((0,0), (1,1), (2,2.1)\)

SVD \(\{0,0\}, \{1,1\},\{2,2.1\}\)

\[
X = \begin{pmatrix}
0 & 0 \\
1 & 1 \\
2 & 2.1 \\
\end{pmatrix}
\]

\((3\times2\ \text{matrix})\)

\[
V^T = \begin{pmatrix}
0.693 & 0.721 \\
0.721 & -0.693 \\
\end{pmatrix}
\]

\((2\times2\ \text{matrix})\)

\[
X = U \Sigma V^T
\]

\[
U = \begin{pmatrix}
\sim 0 & \sim 0 & 1 \\
0.438 & 0.899 & \sim 0 \\
0.899 & -0.438 & \sim 0 \\
\end{pmatrix}
\]

\((3\times3\ \text{matrix})\)

\[
\Sigma = \begin{pmatrix}
3.226 & 0 \\
0 & 0.031 \\
0 & 0 \\
\end{pmatrix}
\]

\((3\times2\ \text{matrix})\)
SVD approximates the exact result

\[ X = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2.1 \end{pmatrix} = u_1 \sigma_1 v_1^T + u_2 \sigma_2 v_2^T \]

\[ u_1 = \begin{pmatrix} 0 \\ -0.438 \\ -0.721 \end{pmatrix}, \quad \sigma_1 = 3.226, \quad v_1^T = \begin{pmatrix} -0.693 \\ -0.721 \end{pmatrix} \]

\[ X = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 0 & 0 \end{pmatrix}, \quad X' = u_1 \sigma_1 v_1^T = \begin{pmatrix} 0 & 0 \\ 0.98 & 1.017 \\ 2.01 & 2.08 \end{pmatrix} \]

\[ X' \sim X \text{ because the projection is approximate} \]
(continued) Projection along PCs

\[ XV = U\Sigma V^T V = U\Sigma \]

(3x2 matrix) \[ X = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 2 & 2.1 \end{pmatrix} \]

(2x2 matrix) \[ V = \begin{pmatrix} -0.693 & -0.721 \\ -0.721 & 0.693 \end{pmatrix} \]

(3x2 matrix) \[ X.V = U.\Sigma = \begin{pmatrix} 0 & 0 \\ 1.414 & -0.028 \\ 2.9 & 0.0133 \end{pmatrix} \]

Projection along PC1
Outline

1. Why do we need reduction in data dimension
2. Theory of Principle Component Analysis
3. Applications of Principle Component Analysis
4. Conclusions
Principle Component Analysis for classification

If you like this book, you will also like that book (because you belong to the same category)
Image Transmission by Principle Component Analysis

\[ X_{1000 \times 500} = u_1 \sigma_1 v_1^T + u_2 \sigma_2 v_2^T + \ldots \]

\[ 1 \times 1 \]
\[ 1 \times 500 \]
\[ 1000 \times 1 \]
%% SVD - Image processing
clearvars;clc;close all;
%%
IMM=imread('MonaLisa.jpg');
IMM=im2double(IMM);
imshow(IMM);

%% Turn it into 2D (grayscale)
IM2D=rgb2gray(IMM);

%% Original in Grayscale
figure(123)
subplot(2,2,1)
imshow(IM2D);
title('Original')

%% SVD Decomposition
[U,S,V]=svd(IM2D);
%% Optional:Plot the diagonals
%figure(234)
%semilogy(diag(S))

%% Keep the first 50 dimensions.
NR = 50; AR50=zeros(size(IM2D));
for ii=1:NR
  AR50=AR50+U(:,ii)*S(ii,ii)*V(:,ii)';
end

%% Keep the first 15 dimensions.
NR = 15; AR15=zeros(size(IM2D));
for ii=1:NR
  AR15=AR15+U(:,ii)*S(ii,ii)*V(:,ii)';
end
Conclusions

1. PCR is a powerful tool to classify multi-dimensional data (e.g. postal codes in handwritten envelopes).

2. PCR decomposition by SVD provides both the rotated axes ($V$) and the projection on the rotated axes ($U \Sigma$). Each column of $(U \Sigma)$ is the projection on that principal component.

3. If there are 100 dimensions and 5 key distinguishing features, then top five singular values may not align with the top five features. One should keep approximately 20 to preserve the top 5 features.

4. The desired by accuracy is obtained by choosing a $k$ such that $p = r_k = \frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{N} \lambda_i^2}$.

5. Other techniques (e.g. Fisher linear discriminators) which finds the direction of the line that best separates two classes may be more accurate or efficient. For example, in Facial recognition, the PCA eigenvalues are called eigenfaces, while that from Fisher LDA is called Fisher’s faces.
Review Questions

1. What is “singular” about singular value decomposition?
2. What is the physical meaning of U and V?
3. How many Principal Components should we need to keep? How do you quantify it?
4. What are the disadvantages of SVD-based classification? In what ways is machine-learning better?
5. What other methods of classification do we have?
6. What applications do we have SVD other than classification (e.g. data compression, etc.)?
7. Taken from your daily experience, Give several examples where SVD classification can be useful.
8. Can you do SVD with Excel? What about Wolfram alpha?
A slightly different approach that also reduces the number of experiments greatly is based on the response surface approach. It uses Newton-like algorithm to find the peaks/valleys of the response surface, see R. H. Myers and D.C. Montgomery, “Response Surface Methodology”, Wiley Interscience, 2002. This book discusses design of experiment in great detail.

For general reference see
