Unsupervised Learning II
Feature Extraction

Feature Extraction Techniques

Unsupervised methods can also be used to find features which can be useful for categorization. There are unsupervised methods that represent a form of smart feature extraction.

- Transforming the input data into the set of features still describing the data with sufficient accuracy
- In pattern recognition and image processing, feature extraction is a special form of dimensionality reduction.

What is feature reduction?

- Most machine learning and data mining techniques may not be effective for high-dimensional data
- When the input data to an algorithm is too large to be processed and it is suspected to be redundant (much data, but not much information)
- Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which overfits the training sample and generalizes poorly to new samples
- The important dimension may be small.
  - For example, the number of genes responsible for a certain type of disease may be small.

Why feature reduction?

\[ G^T \in \mathbb{R}^{d_p \times p}, \quad X \in \mathbb{R}^d, \quad G \in \mathbb{R}^{p \times d} : X \rightarrow Y = G^T X \in \mathbb{R}^d \]
Why feature reduction?

- **Visualization**: projection of high-dimensional data onto 2D or 3D.
- **Data compression**: efficient storage and retrieval.
- **Noise removal**: positive effect on query accuracy.

Feature reduction versus feature selection

- **Feature reduction**
  - All original features are used
  - The transformed features are linear combinations of the original features.

- **Feature selection**
  - Only a subset of the original features are used.

- **Continuous versus discrete**

Application of feature reduction

- Face recognition
- Handwritten digit recognition
- Text mining
- Image retrieval
- Microarray data analysis
- Protein classification

Algorithms

- Feature Extraction Techniques
  - Principal component analysis
  - Singular value decomposition
  - Non-negative matrix factorization
  - Independent component analysis
What is Principal Component Analysis?

- Principal component analysis (PCA)
  - Reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables
  - Retains most of the sample's information.
  - Useful for the compression and classification of data.

- By information we mean the variation present in the sample, given by the correlations between the original variables.
  - The new variables, called principal components (PCs), are uncorrelated, and are ordered by the fraction of the total information each retains.

Principal Components Analysis (PCA)

- Principle
  - Linear projection method to reduce the number of parameters
  - Transfer a set of correlated variables into a new set of uncorrelated variables
  - Map the data into a space of lower dimensionality
  - Form of unsupervised learning

- Properties
  - It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables
  - New axes are orthogonal and represent the directions with maximum variability

Background Mathematics

- Linear Algebra
- Calculus
- Probability and Computing
Consider the matrix
\[
P = \begin{bmatrix} 1 & -1 & 2 \\ 1 & 2 & 3 \\ 13 & 1 & 2 \end{bmatrix}
\]
we have Determinants of \( P \) \( \det(P) \neq 0 \). So this matrix is invertible. Easy calculations give
\[
P^{-1} = \begin{bmatrix} -7 & 24 & 9 \\ -27 & 24 & 9 \\ 32 & 12 & 8 \end{bmatrix}
\]
Next we evaluate the matrix \( P^{-1}AP \).
\[
P^{-1}AP = \begin{bmatrix} -7 & 0 & -7 \\ -27 & 24 & 9 \\ 32 & 12 & 8 \end{bmatrix}
\]
In other words, we have
\[
P^{-1}AP = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 3 \end{bmatrix}
\]
Eigenvalues & eigenvectors
Consider the matrix \( P \) for which the columns are \( C_1, C_2 \), and \( C_3 \), i.e.,
\[
P = \begin{bmatrix} 1 & -1 & 2 \\ 1 & 2 & 3 \\ 13 & 1 & 2 \end{bmatrix}
\]
In other words, we have
\[
AC = \lambda C, \quad \lambda = 0, 4, 9
\]
0, 4 and 3 are eigenvalues of \( A \), \( C_1, C_2 \) and \( C_3 \) are eigenvectors
\[
AC = \lambda C
\]
Example. Consider the matrix
\[
A = \begin{bmatrix} 1 & 2 & 1 \\ 6 & -1 & 0 \\ -1 & -2 & -1 \end{bmatrix}
\]
Consider the three column matrices
\[
C_1 = \begin{bmatrix} 1 \\ 6 \\ 1 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad C_3 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}
\]
We have
\[
AC_1 = 0C_1, \quad AC_2 = -4C_2, \quad AC_3 = 3C_3
\]
In other words, we have
\[
AC_1 = 0C_1, \quad AC_2 = -4C_2, \quad AC_3 = 3C_3
\]
0, 4 and 3 are eigenvalues of \( A \), \( C_1, C_2 \), and \( C_3 \) are eigenvectors
\[
AC = \lambda C
\]
Eigenvalues & eigenvectors
Definition. Let \( A \) be a square matrix. A non-zero vector \( C \) is called an eigenvector of \( A \) if and only if there exists a number (real or complex) \( \lambda \) such that
\[
AC = \lambda C
\]
If such a number \( \lambda \) exists, it is called an eigenvalue of \( A \). The vector \( C \) is called an eigenvector associated to the eigenvalue \( \lambda \).
Remark. The eigenvector \( C \) must be non-zero since we have
\[
A0 = 0 = \lambda 0 \quad (0 \text{ is a zero vector})
\]
for any number \( \lambda \).
Eigenvalues & eigenvectors

Example. Consider the matrix

\[
A = \begin{bmatrix}
1 & 2 & 1 \\
6 & -1 & 0 \\
-1 & -2 & -1
\end{bmatrix}
\]

We have seen that

\[
AC_1 = 0C_1, \quad AC_2 = 4C_2, \quad AC_3 = 3C_3,
\]

where

\[
C_1 = \begin{bmatrix}
1 \\
6 \\
13
\end{bmatrix}, \quad C_2 = \begin{bmatrix}
-1 \\
2 \\
1
\end{bmatrix}, \quad C_3 = \begin{bmatrix}
2 \\
3 \\
-2
\end{bmatrix}.
\]

So \(C_1\) is an eigenvector of \(A\) associated to the eigenvalue 0, \(C_2\) is an eigenvector of \(A\) associated to the eigenvalue -4 while \(C_3\) is an eigenvector of \(A\) associated to the eigenvalue 3.

Determinants

Determinant of order 2

\[
\begin{vmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}
\]

Example: Evaluate the determinant:

\[
\begin{vmatrix}
1 & 2 \\
3 & 4
\end{vmatrix} = 1 \times 4 - 2 \times 3 = -2
\]

Eigenvalues & eigenvectors

For a square matrix \(A\) of order \(n\), the number \(\lambda\) is an eigenvalue if and only if there exists a non-zero vector \(C\) such that

\[
AC = \lambda C
\]

Using the matrix multiplication properties, we obtain

\[
(A - \lambda I)C = 0
\]

We also know that this system has one solution if and only if the matrix coefficient is invertible, i.e.

\[
\det(A - \lambda I) \neq 0
\]

Since the zero vector is a solution and \(C\) is not the zero vector, then we must have

\[
\det(A - \lambda I) = 0.
\]

Example. Consider the matrix

\[
A = \begin{bmatrix}
1 & -2 \\
-2 & 0
\end{bmatrix}
\]

The equation \(\det(A - \lambda I) = 0\) translates into

\[
\begin{vmatrix}
\lambda - 1 & -2 \\
-2 & \lambda - 0
\end{vmatrix} = (\lambda - 1)(\lambda - 0) - 4 = 0
\]

which is equivalent to the quadratic equation

\[
\lambda^2 - \lambda - 4 = 0
\]

Solving this equation leads to (use quadratic formula)

\[
\lambda = \frac{1 \pm \sqrt{17}}{2}, \quad \text{and} \quad \lambda = \frac{1 - \sqrt{17}}{2}
\]

In other words, the matrix \(A\) has only two eigenvalues.
Eigenvalues & eigenvectors

In general, for a square matrix $A$ of order $n$, the equation

$$\det(A - \lambda I_n) = 0.$$  

will give the eigenvalues of $A$.

It is a polynomial function in $\lambda$ of degree $n$. Therefore this equation will not have more than $n$ roots or solutions.

So a square matrix $A$ of order $n$ will not have more than $n$ eigenvalues.

Eigenvalues & eigenvectors

Example. Consider the diagonal matrix

$$D = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}$$

Its characteristic polynomial is

$$\det(D - \lambda I_3) = (a - \lambda)(b - \lambda)(c - \lambda).$$

So the eigenvalues of $D$ are $a$, $b$, and $c$, i.e. the entries on the diagonal.

Computation of Eigenvectors

Let $A$ be a square matrix of order $n$ and $\lambda$ one of its eigenvalues.

Let $X$ be an eigenvector of $A$ associated to $\lambda$. We must have

$$AX = \lambda X \quad \text{or} \quad (A - \lambda I_n)X = 0.$$  

This is a linear system for which the matrix coefficient is $A - \lambda I_n$.

Since the zero-vector is a solution, the system is consistent.

Remark. Note that if $X$ is a vector which satisfies $AX = \lambda X$, then the vector $Y = cX$ (for any arbitrary number $c$) satisfies the same equation, i.e. $AY = \lambda Y$.

In other words, if we know that $X$ is an eigenvector, then $cX$ is also an eigenvector associated to the same eigenvalue.

Eigenvalues & eigenvectors

Computation of Eigenvectors

Example. Consider the matrix

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 6 & -1 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

First we look for the eigenvalues of $A$. These are given by the characteristic equation

$$\det(A - \lambda I) = 0.$$  

If we develop this determinant using the third column, we obtain

$$\begin{vmatrix} 1 - \lambda & 2 & 1 \\ 6 & -1 - \lambda & 0 \\ -1 & -2 & -1 - \lambda \end{vmatrix} = 0.$$  

By algebraic manipulations, we get

$$-\lambda(\lambda + 4)(\lambda - 3) = 0$$

which implies that the eigenvalues of $A$ are $0$, $-4$, and $3$.  

Eigenvalues & eigenvectors
Computation of Eigenvectors

1. Case λ=0: The associated eigenvectors are given by the linear system
   \[ AX = [0] \]
   which may be rewritten by
   \[
   \begin{align*}
   x + 2y + z &= 0 \\
   6x - y &= 0 \\
   -x - 2y - z &= 0
   \end{align*}
   \]
   The third equation is identical to the first. From the second equation, we have \( y = 6x \), so the first equation reduces to \( 13x + z = 0 \).
   So this system is equivalent to
   \[
   \begin{align*}
   y &= 6x \\
   z &= -13x
   \end{align*}
   \]

Computation of Eigenvectors

2. Case λ=−4: The associated eigenvectors are given by the linear system
   \[ AX = −4X \]
   or
   \[ (A + 4I)x = 0 \]
   which may be rewritten by
   \[
   \begin{align*}
   5x + 2y + z &= 0 \\
   6x + 3y &= 0 \\
   -x - 2y + 3z &= 0
   \end{align*}
   \]
   We use elementary operations to solve it.
   First we consider the augmented matrix \( [A + 4I, 0] \)
   \[
   \begin{pmatrix}
   5 & 2 & 1 & 0 \\
   6 & 3 & 0 & 0 \\
   -1 & -2 & 3 & 0
   \end{pmatrix}
   \]
   First we interchange the first row to the end
   \[
   \begin{pmatrix}
   -1 & -2 & 3 & 0 \\
   5 & 2 & 1 & 0 \\
   6 & 3 & 0 & 0
   \end{pmatrix}
   \]
   Next, we use the first row to eliminate the 5 and 6 on the first column. We obtain
   \[
   \begin{pmatrix}
   -1 & -2 & 3 & 0 \\
   0 & -8 & 16 & 0 \\
   0 & -9 & 18 & 0
   \end{pmatrix}
   \]

Computation of Eigenvectors

So the unknown vector \( X \) is given by
\[
X = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 6 \\ -13 \end{pmatrix}
\]
Therefore, any eigenvector \( X \) of \( A \) associated to the eigenvalue 0 is given by
\[
X = c \begin{pmatrix} 1 \\ 6 \\ -13 \end{pmatrix}
\]
where \( c \) is an arbitrary number.
Computation of Eigenvectors

If we cancel the 8 and 9 from the second and third row, we obtain

\[
\begin{pmatrix}
-1 & -2 & 3 & 0 \\
0 & -1 & 2 & 0 \\
0 & -1 & 2 & 0
\end{pmatrix}
\]

Finally, we subtract the second row from the third to get

\[
\begin{pmatrix}
-1 & -2 & 3 & 0 \\
0 & -1 & 2 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\]

Computation of Eigenvectors

Next, we set \( z = c \). From the second row, we get

\[ y = 2z = 2c \] 

The first row will imply \( x = -2y + 3z = -c \). Hence

\[
X = \begin{pmatrix}
x \\
y \\
z \\
c
\end{pmatrix} = \begin{pmatrix}
-1 \\
2 \\
1 \\
c
\end{pmatrix}
\]

Therefore, any eigenvector \( X \) of \( A \) associated to the eigenvalue \(-4\) is given by

\[
X = c \begin{pmatrix}
-1 \\
2 \\
1
\end{pmatrix}
\]

where \( c \) is an arbitrary number.

Computation of Eigenvectors

Case \( \lambda = 3 \): Using similar ideas as the one described above, one may easily show that any eigenvector \( X \) of \( A \) associated to the eigenvalue 3 is given by

\[
X = \begin{pmatrix}
2 \\
3 \\
2
\end{pmatrix}
\]

where \( c \) is an arbitrary number.

Computation of Eigenvectors

Summary: Let \( A \) be a square matrix. Assume \( \lambda \) is an eigenvalue of \( A \). In order to find the associated eigenvectors, we do the following steps:

1. Write down the associated linear system

\[ AX = \lambda X \]

or \( (A - \lambda I)X = 0 \)

2. Solve the system.

3. Rewrite the unknown vector \( X \) as a linear combination of known vectors.
**Why Eigenvectors and Eigenvalues**

An eigenvector of a square matrix is a non-zero vector that, when multiplied by the matrix, yields a vector that differs from the original at most by a multiplicative scalar. The scalar is represented by its eigenvalue.

In this shear mapping the red arrow changes direction but the blue arrow does not. The blue arrow is an eigenvector, and since its length is unchanged its eigenvalue is 1.

**Principal Components Analysis (PCA)**

- **Principle**
  - Linear projection method to reduce the number of parameters
  - Transfer a set of correlated variables into a new set of uncorrelated variables
  - Map the data into a space of lower dimensionality
  - Form of unsupervised learning

- **Properties**
  - It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables
  - New axes are orthogonal and represent the directions with maximum variability

**PCs and Variance**

- The first PC retains the greatest amount of variation in the sample
- The $k$th PC retains the $k$th greatest fraction of the variation in the sample
- The $k$th largest eigenvalue of the correlation matrix $C$ is the variance in the sample along the $k$th PC

**Dimensionality Reduction**

Can ignore the components of lesser significance.

You do lose some information, but if the eigenvalues are small, you don’t lose much.

- $n$ dimensions in original data
- calculate $n$ eigenvectors and eigenvalues
- choose only the first $p$ eigenvectors, based on their eigenvalues
- final data set has only $p$ dimensions
PCA Example –STEP 1

- Subtract the mean from each of the data dimensions. This produces a data set whose mean is zero. Subtracting the mean makes variance and covariance calculation easier by simplifying their equations. The variance and co-variance values are not affected by the mean value.

<table>
<thead>
<tr>
<th>DATA:</th>
<th>ZERO MEAN DATA:</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>2.5</td>
<td>2.4</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td>2.2</td>
<td>2.9</td>
</tr>
<tr>
<td>1.9</td>
<td>2.2</td>
</tr>
<tr>
<td>3.1</td>
<td>3.0</td>
</tr>
<tr>
<td>2.3</td>
<td>2.7</td>
</tr>
<tr>
<td>2</td>
<td>1.6</td>
</tr>
<tr>
<td>1</td>
<td>1.1</td>
</tr>
<tr>
<td>1.5</td>
<td>1.6</td>
</tr>
<tr>
<td>1.1</td>
<td>0.9</td>
</tr>
</tbody>
</table>

PCA Example –STEP 2

- Calculate the covariance matrix

\[
\text{cov} = \begin{pmatrix}
.616555556 & .615444444 \\
.615444444 & .716555556
\end{pmatrix}
\]

Variance measures how far a set of numbers is spread out

\[\text{Var}(X) = E[(X - \mu)^2] \]

Covariance provides a measure of the strength of the correlation between two or more sets of random variates. The covariance for two random variates \( X \) and \( Y \), each with sample size, is defined by the expectation value

\[\sigma_{x,y} = E[(X - E[X])(Y - E[Y])]\]
PCA Example –STEP 3

- Calculate the eigenvectors and eigenvalues of the covariance matrix

$$\text{eigenvalues} = \begin{pmatrix} 0.04908340 \\ 1.28402771 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} -0.73517866 & -0.677873399 \\ 0.67787340 & -0.735178656 \end{pmatrix}$$

- Eigenvectors are plotted as diagonal dotted lines on the plot.
- Note they are perpendicular to each other.
- Note one of the eigenvectors goes through the middle of the points, like drawing a line of best fit.
- The second eigenvector gives us the other, less important, pattern in the data, that all the points follow the main line, but are off to the side of the main line by some amount.

PCA Example –STEP 4

- Reduce dimensionality and form feature vector
  
  the eigenvector with the highest eigenvalue is the principle component of the data set.

In our example, the eigenvector with the largest eigenvalue was the one that pointed down the middle of the data.

Once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives you the components in order of significance.

- Now, if you like, you can decide to ignore the components of lesser significance.

  You do lose some information, but if the eigenvalues are small, you don’t lose much

  - n dimensions in your data
  - calculate n eigenvectors and eigenvalues
  - choose only the first p eigenvectors
  - final data set has only p dimensions.
PCA Example –STEP 4

• Feature Vector

FeatureVector = (eig_1 eig_2 eig_3 … eig_n)

We can either form a feature vector with both of the eigenvectors:

\[
\begin{pmatrix}
-0.677873399 & -0.735178656 \\
-0.735178656 & 0.677873399
\end{pmatrix}
\]

or, we can choose to leave out the smaller, less significant component and only have a single column:

\[
\begin{pmatrix}
-0.677873399 \\
-0.735178656
\end{pmatrix}
\]

PCA Example –STEP 5

• Deriving the new data

FinalData = RowFeatureVector x RowZeroMeanData

RowFeatureVector is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvector at the top.

RowZeroMeanData is the mean-adjusted data transposed, i.e., the data items are in each column, with each row holding a separate dimension.

FinalData transpose: dimensions along columns

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.827970186</td>
<td>-0.175115307</td>
</tr>
<tr>
<td>-1.77758033</td>
<td>0.142857227</td>
</tr>
<tr>
<td>-0.992197494</td>
<td>0.384374989</td>
</tr>
<tr>
<td>-0.274210416</td>
<td>0.130417207</td>
</tr>
<tr>
<td>-1.67580142</td>
<td>-0.209498461</td>
</tr>
<tr>
<td>-0.912949103</td>
<td>0.175282444</td>
</tr>
<tr>
<td>-0.0991094375</td>
<td>-0.349824698</td>
</tr>
<tr>
<td>1.14457216</td>
<td>0.0464172582</td>
</tr>
<tr>
<td>0.438046137</td>
<td>0.0177646297</td>
</tr>
<tr>
<td>1.22382056</td>
<td>-0.162675287</td>
</tr>
</tbody>
</table>


Figure 5.3: The table of data by applying the PCA analysis using both eigenvectors, and a plot of the new data points.
PCA Algorithm

- Get some data
- Subtract the mean
- Calculate the covariance matrix
- Calculate the eigenvectors and eigenvalues of the covariance matrix
- Choosing components and forming a feature vector
- Deriving the new data set

Why PCA?

- Maximum variance theory
  - In general, variance for noise data should be low and signal should be high. A common measure is the signal-to-noise ratio (SNR). A high SNR indicates high precision data.

Projection

\[ \mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos(\theta) \]

The dot product is also related to the angle between the two vectors but it doesn’t tell us the angle.

\[ \mathbf{x}^T \mathbf{u} \] is the length from blue node to origin of coordinates. The mean for the \( \mathbf{x}^T \mathbf{u} \) is zero.

\[ \text{var}(\mathbf{x}^T \mathbf{u}) = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}^T \mathbf{u})^2 = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}^T \mathbf{x}^T \mathbf{u} \]

\[ = \mathbf{u}^T \left( \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}^T \mathbf{x}^T \mathbf{u} \right) \mathbf{u} \]

Covariance Matrix since the mean is zero

\[ \text{Cov} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}^T \mathbf{x}^T \mathbf{u} \]

\[ = \mathbf{u}^T \left( \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}^T \mathbf{x}^T \mathbf{u} \right) \mathbf{u} \]

\[ \mathbf{u} = \mathbf{u} \mathbf{u} \text{ Cov} \mathbf{u} = \text{Cov} \mathbf{u} \]

Therefore, \( \mathbf{Cov} \mathbf{u} = \mathbf{u} \)

We get it. \( \lambda \) is the eigenvalue of matrix \( \mathbf{Cov} \mathbf{u} \) is the Eigenvectors. The goal of PCA is to find an \( \mathbf{u} \) where the variance of all projection points is maximum.