Recitation

10-701/15-781, Fall 2008

PCA

By Hanghang Tong
(htong@cs)

Some Parts of the slides are from previous years' recitation and lecture notes
Roadmap

- PCA: Maximize projected variance
- PCA: Minimize reconstruction error
- PCA vs. SVD
- PCA vs. Feature Selection
PCA: Settings

- We have N data points in D dimension space, we want to project them into some low-dimensional space.

Here, N=5; D=2; and d=1

Q: How shall we select \( u \) (the projection direction)?
PCA: Maximize the variance

- Q: How to choose (u)?
- A: Maximize the variance among red points?
Maximize Variance: (d=1)

Projected data

Original data

\( \bar{X}_i = u^T X_i \ (i = 1, \ldots, 5) \)

\( \overline{X} = \frac{\tilde{X}_1 + \tilde{X}_2 + \tilde{X}_3 + \tilde{X}_4 + \tilde{X}_5}{5} \)

\( = u^T \left( X_1 + X_2 + X_3 + X_4 + X_5 \right) \frac{1}{5} \)

\( = u^T \overline{X} \)

the variance among red points =

\[ \frac{(\tilde{X}_1 - \overline{X})^2 + (\tilde{X}_2 - \overline{X})^2 + (\tilde{X}_3 - \overline{X})^2 + (\tilde{X}_4 - \overline{X})^2 + (\tilde{X}_5 - \overline{X})^2}{5} \]
Maximize Variance: (d=1)

- The variance $J$ for the projected data:

$$J = \frac{1}{N} \sum_{n=1}^{N} \left( u_1^T x_n - u_1^T \bar{x} \right)^2$$

$$= \frac{1}{N} \sum_{n=1}^{N} u_1^T (x_n - \bar{x})(x_n - \bar{x})^T u_1$$

$$= u_1^T \left( \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T \right) u_1$$

$$= u_1^T S u_1$$

Sample Variance Matrix $D \times D$
Maximize Variance: (d=1)

- Optimization Problem Formulation:

\[
\begin{align*}
\max J &= u_1^T S u_1 \\
\text{subject to: } u_1^T u_1 &= 1
\end{align*}
\]

PC has the unit length

Q: How to Solve it?
Maximize Variance: \((d=1)\)

- **How to Solve**

\[
\begin{align*}
\max J &= u_1^T S u_1 \\
\text{subject to: } u_1^T u_1 &= 1
\end{align*}
\]

- **Step 1:** Write down the Lagrange function

\[
J' = u_1^T S u_1 - \lambda (u_1^T u_1 - 1)
\]

- **Step 2:**

Set \(\frac{\partial J'}{\partial u_1} = S u_1 - \lambda u_1\) to be zero

- **We have**

\[
S u_1 = \lambda u_1
\]
Maximize Variance: (d=1)

- We already know
  \[ Su_1 = \lambda u_1 \]

- We want to show \( \lambda \) must be the 1st eigen-value of \( S \)
Maximize Variance: \((d>1)\)

- Fact: \(d\) PCs are the first \(d\) eigen-vectors of \(S\)

Why? By induction

- Step 0: Assume that this fact holds for \(d=M\)
- Step 1: for \(d=M+1\), the variance \(J\) of the projected data is
  \[ J = \sum_{j=1}^{M+1} u_j^T S u_j \]
- Step 2: Write down the opt. formulation

\[
\begin{align*}
\max \tilde{J} & = u_{M+1}^T S u_{M+1} \\
\text{subject to} : & \quad u_j^T u_{M+1} = 0 (j = 1, \ldots, M); \quad u_{M+1}^T u_{M+1} = 1
\end{align*}
\]

PCs are orthogonal w/ each other

PCs have unit length

The first \(M\) PCs are fixed
Maximize Variance: (d>1)

- Step 3: Write down the Lagrange function

\[ J' = u_{M+1}^T S u_{M+1} - \sum_{j=1}^{M} \eta_j u_j^T u_{M+1} - \lambda (u_{M+1}^T u_{M+1} - 1) \]

- Step 4:

- We have

\[ Su_{M+1} = \sum_{j=1}^{M} \eta_j u_j + \lambda u_{M+1} \]

- PCs are orthogonal w/ each other
- PCs have unit length
Maximize Variance: (d>1)

- We have

\[ S u_{M+1} = \sum_{j=1}^{M} \eta_j u_j + \lambda u_{M+1} \]

1. Want to show \( \eta_j = 0 \)

2. Want to relate \( \lambda \) to \( \tilde{\lambda} \) (do it on your own)

- We want to show \( u_{M+1} \) is the (M+1)th eigenvector of \( S \)
Maximize the variance vs. Minimize the reconstruction error

- Q: How to choose (u)?
  - Choice 1: Maximize the variance of projected data
    - Make red points as scattered as possible
  - Choice 2: Minimize the reconstruction error
    - Make the green curves as small as possible
- Fact: Choice 1 = Choice 2
PCA: Minimize Reconstruction Error (Assuming zero mean)

- We have $N$ data points in $D$ dimension space

$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}_{N \times D}$

$\tilde{X} = \begin{bmatrix} \tilde{X}_1 \\ \tilde{X}_2 \\ \vdots \\ \tilde{X}_N \end{bmatrix}_{N \times d}$

$U = \begin{bmatrix} u_1 \\ \vdots \\ u_d \end{bmatrix}_{d \times D}$

Goal: make $\left\| X - \tilde{X} \cdot U \right\|_F$ as small as possible
PCA vs. SVD (assuming zero mean)

Goal: make \( \|X - \hat{X} \cdot U\|_F \) as small as possible

- It turns out...
PCA vs. Feature Selection

- We have $N$ data points in $D$ dimension space

\[ X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} F_1 \\ \vdots \\ F_d \end{bmatrix} \]

- In PCA
  - Un-supervised
  - Generate a few new features (linear combination of original columns)

- Feature Selection
  - Supervised (with class labels or regression output)
  - Select original features (columns)
End