PCA Review   Author: Shaohong Luo

Section 1. Some important Linear Algebra background for PCA

- **Some useful conclusion**

For a matrix $A \in \mathbb{R}^{m \times n}$

① A^TA and AA^T is symmetric matrix

② A^TA and AA^T can be diagonalizable and get an orthonormal eigenvector.

③ $\operatorname{rank}(A) = \operatorname{rank}(A^T) = \operatorname{rank}(A^TA) = \operatorname{rank}(A^TA)$

④ A^TA is positive semi-definite. If all the column of A is independent, then A^TA is positive definitive.

⑤ A^TA and AA^T have the same non-zero eigenvalues. The number of non-zero eigenvalues is equal to $\operatorname{rank}(A)$.

- **Spectral theorem**

For a symmetric matrix, $A = UDU^T$. where $UU^T = U^TU = I$

- **SVD decomposition**

① For a matrix $A_{m \times n}$. $A = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T$, where $U^TU = UU^T = I$, $V^TV = VV^T = I$, Σ is diagonal. $\Sigma = \begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_r & \\ & & & & 0 \end{pmatrix}$, $\sigma_1 > \sigma_2 > \dots > \sigma_r$

② Question: How to compute SVD?

1) Right singular vector: Find eigenvectors of A^TA. This give V matrix

2) Singular Value: Find eigenvalues of A^TA. AA^T has $\operatorname{rank}(A)$ eigenvalues and A^TA / AA^T has the same non-zero eigenvalue. (Conclusion ⑤)

$$\sigma_j = \sqrt{\lambda_j} \quad 1 \le j \le r$$
$$= 0 \quad r+1 \le j \le n$$

3) Left singular vector: $u_j = \frac{1}{\sigma_j} A v_j \quad 1 \le j \le r$

- **Truncated SVD**

Ignore some of the small singular vector.

$\tilde{A} = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$ (Only use the Top 2 largest σ)

• Matrix norm and Eckart-Young Theorem

① Frobenius norm of A is $\|A\|_F^2 = \sum_{ij} a_{ij}^2 = \text{trace}(A^TA) = \sum_{i=1}^{r} \sigma_i^2$

② ℓ_2 norm of a matrix $\|A\|_2 = \max\limits_{1\le j\le n} \sigma_{max} = \sigma_1$

③ Eckart-Young Theorem:

$A^{m\times n}$ is a rank r matrix, $B^{m\times n}$ is a rank k matrix where $k \le r$

Define $\hat{A}_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$, then $\begin{cases} \|A - \hat{A}_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2 \\ \|A - \hat{A}_k\|_2 = \sigma_{k+1} \end{cases}$

④ Question: How to decide k in low rank approximation?

where $\hat{A}_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$

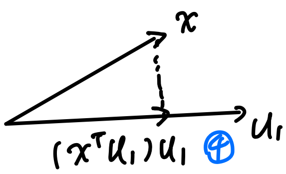
Answer: $\dfrac{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_k^2}{\sigma_{k+1}^2 + \sigma_{k+2}^2 \cdots + \sigma_n^2} >$ threshold (0.95 / 0.90 ...)

Section 2. PCA

• 2-prospectives

• Maximum variance of projection

Project onto u_1: Find a direction u_1 such that the variance of the projection of the data onto u_1 is maximized.



$$\arg\max_{u_1} \frac{1}{n} \sum_{i=1}^{n} \left[((x-\bar{x})^T u_1) u_1\right]^2, \quad u^Tu = 1$$

$$\arg\max_{u_1} \frac{1}{n} \sum_{i=1}^{n} \left[(x^Tu_1) u_1\right]^2 \quad ①$$

$$\Rightarrow \arg\max_{u_1} \frac{1}{n} \sum_{i=1}^{n} (u_1^T x)^2 \quad ②$$

$$\Rightarrow \arg\max_{u_1} \frac{1}{n} \sum_{i=1}^{n} (u_1^T x)(u_1^T x)^T$$

$$\Rightarrow \arg\max_{u_1} \frac{1}{n} \sum_{i=1}^{n} u_1^T x x^T u_1$$

$$\Rightarrow \arg\max_{u_1} u_1^T \left(\frac{1}{n} \sum_{i=1}^{n} x x^T\right) u_1$$

Note that $\frac{1}{n} \sum_{i=1}^{n} x x^T$ is a covariance matrix.

define $S = \frac{1}{n} \sum_{i=1}^{n} x x^T$

Then: $\arg\max\limits_{u_1} u_1^T S u_1$

The optimization problem above can be written as

$$\begin{cases} \arg\max_{u_1} u_1^T S u_1 \\ u_1^T u_1 = 1 \end{cases}$$

Apply Lagrange multiplier:

$$\ell(u_1, \lambda) = u_1^T S u_1 - \lambda(u_1^T u_1 - 1) = 0$$

$$\frac{d\,\ell(u_1,\lambda)}{d\,u_1} = 2 u_1 S - 2\lambda u_1 = 0 \quad ③$$

$$\Rightarrow \underline{u_1 S = \lambda u_1}$$

where λ is the eigenvalue of S and u_1 is the eigenvector of S

Remark

① For convenience, assume x has been centered such that $\bar{x} = 0$

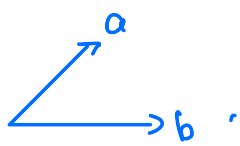
② $(x^T u)$ is a coefficient (constant value), therefore $x^T u = u^T x$

③ Apply the formula of matrix derivation:

$$\begin{cases} \dfrac{\partial x^T A x}{\partial x} = 2Ax \quad (A \text{ is symmetric}) \\ \dfrac{\partial x^T x}{\partial x} = 2x \end{cases}$$

④ vector projection formula:

$$\mathrm{Proj}_{\vec{b}}\,\vec{a} = \frac{\vec{a}\cdot\vec{b}}{\|\vec{b}\|^2}\,\vec{b}$$



, in PCA, $\|\vec{b}\|^2 = 1$

Summary:

According to the procedure above, PCA can be generalized as 4 steps.

① centralize data.

② Form covariance matrix $S = XX^T$

③ Find eigenvectors and eigenvalues of S $\quad u_1 S = \lambda u_1$

④ The larger the eigenvalue $\Leftrightarrow$ the more significant direction

Minimum reconstruction error

Choose k PCs (principal components)

$$\tilde{V} = \begin{pmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_k \\ | & | & & | \end{pmatrix} \qquad k << d$$

Project data point x_i onto the subspace spanned by the first K PCs.

$$\tilde{x}_i = (x_i^T v_1)v_1 + (x_i^T v_2)v_2 + \cdots + (x_i^T v_k)v_k$$

$$= (x_i^T v_1)v_1 + (x_i^T v_2)v_2 + \cdots + (x_i^T v_k)v_k$$

$$= v_1^T v_1 x_i + v_2^T v_2 x_i + \cdots + v_k^T v_k x_i \qquad \text{(Remark ②)}$$

$$= \tilde{V}^T V x_i$$

The PCA reconstruction error (Pearson 1991) is defined as

$$\boxed{E = \frac{1}{n} \sum_{i=1}^{N} \| x_i - \tilde{x}_i \|^2}$$

Given that $x_i = \sum_{j=1}^{d} v_j^T v_j x_i$, $\tilde{x}_i = \sum_{j=1}^{k} v_j^T v_j x_i$

Then,

$$E = \frac{1}{n} \sum_{i=1}^{N} \| \sum_{j=1}^{d} v_j^T v_j x_i - \sum_{j=1}^{k} v_j^T v_j x_i \|^2$$

$$= \frac{1}{n} \sum_{i=1}^{N} \| \Big(\sum_{j=k+1}^{d} v_j^T v_j \Big) x_i \|^2$$

$$= \frac{1}{n} \sum_{i=1}^{N} x_i^T \Big(\sum_{j=k+1}^{d} v_j^T v_j \Big) \Big(\sum_{q=k+1}^{d} v_j^T v_j \Big)^T x_i$$

$$= \frac{1}{n} \sum_{i=1}^{N} \sum_{j=k+1}^{d} (v_j^T x_i)(x_i^T v_j)$$

$$= \frac{1}{n} \sum_{i=1}^{N} \sum_{j=k+1}^{d} v_j^T (x_i x_i^T) v_j$$

$$= \sum_{j=k+1}^{d} v_j^T \underline{\frac{1}{n} \sum_{i=1}^{N} (x_i x_i^T)} \, v_j$$

$$\boxed{E = \sum_{j=k+1}^{d} v_j^T S v_j}$$

According to our intuition, we want to minimize E.

Note that:

$$\underbrace{\sum_{j=1}^{K} v_j^T S v_j}_{\text{First perspective: maximize this term}} + \underbrace{\sum_{j=K+1}^{d} v_j^T S v_j}_{\text{second perspective: minimize this term}} = \sum_{j=1}^{d} v_j^T S v_j$$

Section 3 Kernel PCA

We have n points $x_1, x_2, \cdots x_n$ each lying in R^d

- standard PCA doesn't yield good features on highly non-linear datasets.
- In detail. define ϕ is some non-linear transformation

  $\phi : \mathbb{R}^d \to \mathbb{R}^m$, $m \gg d$
- We imitate PCA procedures in above highly non-linear datasets
  - Assume that $\phi(x_i)$ has been centralized $\frac{1}{N}\sum_{i=1}^{N}\phi(x_i) = 0$
  - Form covariance matrix $S = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\phi(x_i)^T$ ①
  - Find eigs $v_k S = \lambda_k v_k$. $k = 1, 2, \cdots, M$ ②

Combined ①, ②:

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\left(\phi(x_i)^T v_k\right) = \lambda_k v_k \quad ③$$

Note that each v_k is a linear combination of $\phi(x_i)$.

$$\boxed{v_k = \sum_{j=1}^{N} a_{kj}\,\phi(x_j)} \quad ④$$

Combined ③ ④:

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_i)\,\phi(x_i)^T\sum_{j=1}^{N}a_{kj}\,\phi(x_j) = \lambda_k\sum_{i=1}^{N}a_{ki}\,\phi(x_i) \quad ⑤$$

Define $k(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ ⑥

Multiply $\phi(x_\ell)^T$ both sides in ⑤:

$$\frac{1}{N}\sum_{i=1}^{N}\phi(x_\ell)^T\phi(x_i)\,\phi(x_i)^T\sum_{j=1}^{N}a_{kj}\,\phi(x_j) = \lambda_k\sum_{i=1}^{N}\phi(x_\ell)^T a_{ki}\,\phi(x_i) \quad ⑦$$

Rewrite ⑦ using ⑥ :

$$\frac{1}{N}\sum_{i=1}^{N} k(x_l, x_i)\sum_{j=1}^{N} a_{kj}\, k(x_i, x_j) = \lambda_k \sum_{i=1}^{N} a_{ki}\, k(x_l, x_i) \quad ⑧$$

Define $K_{ij} = k(x_i, x_j)$

Rewrite ⑧ :

$$\frac{1}{N}\sum_{i=1}^{N} K_{li}\sum_{j=1}^{N} a_{kj} K_{ij} = \lambda_k \sum_{i=1}^{N} a_{ki} K_{li} \quad ⑨$$

Left side of ⑨ :

$$\frac{1}{N}\sum_{i=1}^{N} K_{li}\sum_{j=1}^{N} a_{kj} K_{ij} = \frac{1}{N}\left(\sum_{i=1}^{N} K_{li}\right)\left(\sum_{j=1}^{N} \underline{K_{ij}}\, a_{kj}\right)$$

$$= \frac{1}{N}\underline{K} \cdot K a_k$$

$$= \frac{1}{N} K^2 a_k$$

i and l cannot be anything.
by using all the values of
i/l, we will get the vector
$K \cdot a_k$

Right side of ⑨ :

$$\lambda_k \sum_{i=1}^{N} a_{ki} K_{li} = \lambda_k \left(\sum_{i=1}^{N} \underline{K_{li}}\, a_{ki}\right)$$

$$= \lambda_k \underline{K} a_k$$

Hence, rewrite ⑨ :

$$\frac{1}{N} K^2 a_k = \lambda_k K a_k \quad ⑩$$

For non-zero eigenvalues,

$$\frac{1}{N} K a_k = \lambda_k a_k$$

$$\boxed{K a_k = N \lambda_k a_k} \quad ⑪$$

Converted into eigens problem!

- Question: Project $\phi(x)$ onto V_k ?

A: $\phi(x) = \underline{(\phi(x)^T V_k)}\, V_k$

$$\phi(x)^T V_k = \phi(x)^T \sum_{j=1}^{N} a_{kj}\, \phi(x_j)$$

$$= \sum_{j=1}^{N} a_{kj}\, \phi(x)^T \phi(x_j)$$

$$= \sum_{j=1}^{N} a_{kj}\, k(x, x_j)$$

• Example:

Give a non-linear mapping: $\phi\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} u_1^2 \\ u_1 u_2 \\ u_2 u_1 \\ u_2^2 \end{pmatrix}$ 4-d

compute $\phi^T(u) \cdot \phi(v)$, given $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$, $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$

Solution:

$$\begin{pmatrix} u_1^2 & u_1 u_2 & u_2 u_1 & u_2^2 \end{pmatrix} \begin{pmatrix} v_1^2 \\ v_1 v_2 \\ v_2 v_1 \\ v_2^2 \end{pmatrix}$$

$$= u_1^2 v_1^2 + 2 u_1 u_2 v_1 v_2 + u_2 v_2^2$$

$$= (u_1 v_1 + u_2 v_2)^2$$

$$= (u^T v)^2$$ $\leftarrow$ no need to form 4-dim vectors to compute $\phi(u)^T \phi(v)$

Note: We don't need to know ϕ is explicitly. All we care is being able to compute the kernel.

Question: What is the kernel ($k(x_i, x_j)$)?

Answer: Give some kernel following:

polynomial kernel: $k(x_1, x_2) = (1 + x_1^T x_2)^m$

Gaussian kernel (RBF): $k(x_1, x_2) = \exp\left(\frac{-\|x_1 - x_2\|^2}{2\sigma^2}\right)$

...

Question: What are the conditions to be a kernel?

Answer: Mercer's condition: $k(x, x')$ is valid kernel function if and only if the kernel matrix is always symmetric positive semi-definite for any given $\{x_1, x_2, \cdots, x_n\}$

Example: Check $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is a "qualified" kernel function.

Solution:

Theorem: if K is positive semi-definite, its quadratic form must ≥ 0

Hence, we just need to check if $y^T K y \geq 0$

$$
\begin{aligned}
y^T K y &= y^T \phi(x_i)^T \phi(x_j)\, y \\
&= \sum_{ij} \phi(x_i)^T \phi(x_j)\, y_i y_j \\
&= \left[\sum_i y_i \phi(x_i)^T\right]^T \sum_j y_j \phi(x_j) \\
&= \Phi^T y^T y \Phi, \quad \text{where } \Phi = \begin{pmatrix} | & & | \\ \phi(x_1) & \cdots & \phi(x_n) \\ | & & | \end{pmatrix} \\
&= (y\Phi)^T (y\Phi) \geq 0
\end{aligned}
$$