Pca No-Lineal y Kernel-PCA

Antonio Sala Piqueras

Identificación de sistemas complejos

Dept. Ing. Sistemas y Automatica (DISA)

Universitat Politècnica de València (UPV)

Presentación

Motivación:
El PCA lineal es la búsqueda de proyecciones ortogonales de unos datos sobre subespacios de poca/mucha varianza. ¿Caso no lineal? ¿Caso mucha info por muestra, pero pocas muestras?

Objetivos:
Comprender la generalización del PCA a caso no lineal y a caso Kernel.

Contenidos:
Planteamiento del problema. Variantes del PCA. Truco Kernel. Centrado y reconstrucción. Conclusiones.

Planteamiento del problema

Tenemos N datos $x_i \in \mathbb{R}^n$, $i = 1, \ldots, N$. También tenemos N variables a precedir $y_i \in \mathbb{R}^m$. Calculamos q regresores (features) $\phi(x)$, $\phi : \mathbb{R}^n \mapsto \mathbb{R}^q$.

Objetivo PCA: obtener el $\theta_{1\times q}$ que minimiza/maximiza/silla $J_0 = \sum_{i=1}^{N} \|\theta\phi(x(i))\|^2$, sujeto a $\|\theta\| = \theta\theta^T = 1$.

Metodología: Se generan matriz de datos $X_{q\times N} := [\phi(x_1) \ \ldots \ \phi(x_N)]$, y se realizan operaciones sobre ella.

Añadiendo mult. Lagrange: $J(\theta) = \theta X X^T \theta^T - \lambda(\theta\theta^T - 1)$.

Dos escenarios:

- Pocas "features", muchos datos ($q < N$) vs. Muchas "features", pocos datos $q \geq N$.

Variantes del PCA

PCA lineal: $\phi(x) = x \in \mathbb{R}^n$. $N >> n = q$. Solución sin Kernel: $\theta =$ autovector asociado al autovalor más pequeño de XX^T; (XX^T) tiene dimensiones $n \times n$.

PCA NO lineal: Supongamos que ampliamos nuestra información al algoritmo con más características (features) $\phi(x) \in \mathbb{R}^q$.

Versión "clásica" con "muchas" muestras $N >> q$:, Entonces $\phi([x_1 \dots x_N])\phi([x_1 \dots x_N])^T$ tiene dimensiones $q \times q$, y hacemos PCA. Detectamos combinaciones de características (posiblemente no lineales) cercanas a cero sobre los datos... o aquéllas (perp. a las anteriores) más lejanas de cero, que diferencian más a unas muestras de otras.

PCA no lineal: pocas muestras

Con "pocas" muestras y q grande, $(q > N)$ (XX^T) tiene dimensiones $q \times q$ pero, dado que X es $q \times N$, el rango de XX^T no puede ser mayor que N... Esos autovectores asociados al espacio nulo de XX^T indican direcciones "no exploradas en las muestras" hay que eliminarlos (regularizar).

Mejor aún, **ni siquiera calcularlos:** hay que evitar formar una matriz "enorme" con un montón de autovalores $(q - N)$ nulos... y evitar ni siquiera tener que construir ϕ enorme.

¡Para eso está el "Kernel trick", claro!

El "truco" del Kernel

$$\min_{\theta} J(\theta) = \theta X X^T \theta^T - \lambda(\theta\theta^T - 1) = \min_{w} J(wX^T)$$

$$J(w) = wX^T(XX^T)Xw^T - \lambda(wX^TXw^T - 1) = wK^2w^T - \lambda(wKw^T - 1)$$

siendo K la matriz de productos escalares

$$K_{N\times N} = X^TX = \begin{pmatrix} \phi^T(x_1)\phi(x_1) & \phi^T(x_1)\phi(x_2) & \cdots & \phi^T(x_1)\phi(x_N) \\ \vdots & \ddots & & \vdots \\ \phi^T(x_N)\phi(x_1) & \phi^T(x_N)\phi(x_2) & \cdots & \phi^T(x_N)\phi(x_N) \end{pmatrix}$$

A la matriz K se le denomina **Kernel matrix**.

Kernel PCA

Deriv resp. w: $0 = K^2 w^T - \lambda K w^T = K(Kw^T - \lambda w^T)$.
O sea, $Kw^T = \lambda w^T$, o (K simétrica) equivalentemente $wK = \lambda w$.

Los componentes principales están asociados a los autovectores $w^{[i]}$ y los autovalores λ_i de K, ordenados por varianza.

Normalizado: La condición $wKw^T = 1$ implica $w^{[i]}\lambda_i(w^{[i]})^T = 1$, esto es $\|w^{[i]}\| = \frac{1}{\sqrt{\lambda_i}}$.

Centrado (1)

Para que tenga interpretación PCA X debe ser una matriz con media de columnas igual a cero. Si tenemos $X = [\phi(x_1) \ \ldots \phi(x_N)]_{q\times N}$, entonces la media es

$$\bar{\phi} := \sum_{i=1}^{N} \phi(x_i) = X \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{N\times 1} \cdot \frac{1}{N}$$

Con lo que la nueva X en variables incrementales es la original menos la media repetida N veces, que denominaremos $\bar{X}$:

$$\bar{X}_{N\times N} := \bar{\phi} \cdot [1 \ \ldots 1]_{1\times N} = X \cdot \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \vdots & & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix}_{N\times N} \cdot \frac{1}{N} = X \cdot [\mathbf{1}_{/N}]_{N\times N}$$

Centrado (2)

Por tanto

$$
\begin{aligned}
K_{PCA} &= (X - \bar{X})^T (X - \bar{X}) \\
&= X^T X - \underbrace{[\mathbf{1}_{/N}] X^T}_{\bar{X}^T} X - X^T \underbrace{X \cdot [\mathbf{1}_{/N}]}_{\bar{X}} + \underbrace{[\mathbf{1}_{/N}] X^T}_{\bar{X}^T} \underbrace{X \cdot [\mathbf{1}_{/N}]}_{\bar{X}} \\
&= K - [\mathbf{1}_{/N}] K - K [\mathbf{1}_{/N}] + [\mathbf{1}_{/N}] K [\mathbf{1}_{/N}]
\end{aligned}
$$

Reconstrucción/ cambio vble.

Dado un punto x, la proyección sobre el componente principal i-ésimo será:

$$\theta^{[i]}\phi(x) = (w^{[i]}X^T) \cdot \phi(x) = w^{[i]}\kappa(X, x)$$

En particular, la proyección de cada punto del conjunto de datos será el elemento correspondiente del vector fila:

$$\eta^{[i]} := (w^{[i]}X^T) \cdot X = w^{[i]}K = w^{[i]}\lambda_i$$

cuya norma es $\|\eta^{[i]}\| = \sqrt{\lambda_i}$, usando que $\|w^{[i]}\| = \frac{1}{\sqrt{\lambda_i}}$.

Matlab: Como `eig` devuelve vect. de norma 1, si `[V,D]=eig(K)`, la transformación K-PCA del conjunto de datos será $X_{PCA} = D^{1/2}V^T$.

Trabajo en linea con Kernels centrados

Si w se ha calculado con Kernels centrados, la descomposición $X_{PCA} = D^{1/2}V^T$ es correcta a partir de $eig(K_{PCA})$.

Si se desea usar Kernel PCA para diagnóstico, clasificación o detección de fallos en linea, con nuevos datos, entonces se debe usar:

$$\theta^{[i]}(\phi(x)-\bar{\phi}) = (w^{[i]}X^T)\cdot(\phi(x)-\bar{\phi}) = w^{[i]}\left(\kappa(X,x) - K\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}_{N\times 1}\cdot\frac{1}{N}\right)$$

El resultado es un vector de proyecciones sobre cada componente que debería compararse con los de los datos de entrenamiento X_{PCA} para determinar si es anormalmente grande.

Conclusiones

- El análisis de componentes principales puede aplicarse, en vez de a los valores propios de la covarianza $\Sigma = (XX^T)_{q\times q}$, a los valores propios de los productos escalares $K = (X^TX)_{N\times N}$
- Como todos los métodos Kernel, recomendado con $q \geq N$, o si se escoge K de literatura sin conocer explícitamente $\phi(x)$.
- Para que las conclusiones sean correctas, K debe ser un Kernel "centrado". Si no lo es, se centra con la fórmula:

$$K_{PCA} = K - [\mathbf{1}_{/N}]K - K[\mathbf{1}_{/N}] + [\mathbf{1}_{/N}]K[\mathbf{1}_{/N}]$$

- El conjunto de datos se proyecta, si `[V,D]=eig(`K_{PCA}`)`, en coordenadas de componentes principales $D^{1/2}V^T$, escogiendo filas que convengan para análisis/reducción de dimensionalidad.