

Cap. 4. Selección y extracción de características

4.1. Extracción

La información que pueden aportar objetos de señales 2D y 1D (Datos), puede ser caracterizada por características las cuales en muchas ocasiones se extraen a partir de un mapeo. El propósito de dicho mapeo es reducir la cantidad de datos a partir de una medida de ciertas "características" o "propiedades". Dichas características serán luego alimentadas hacia un clasificador que evaluará la evidencia presentada y tomará una decisión sobre la clasificación de especies u objetos.

Un "extractor" de características ideal brindará una representación que hará que el trabajo del clasificador sea trivial. De igual manera, un buen clasificador no necesitaría la ayuda de un extractor de características sofisticado.

De manera general, la tarea de extracción de características corresponde más a un problema de mapeo que a un problema de clasificación, y requiere de un conocimiento del dominio.

Un buen extractar de características para clasificar peces podría ser de poco uso para la identificación de huellas dactilares, o bien clasificar células de sangre podría requerir de un proceso de extracción particular.

- ¿Cómo saber cuáles características son las más prometedoras?
- ¿Existen formas de aprender de forma automática que características son las mejores para un clasificador?
- ¿Qué tantas características deberíamos utilizar?

II

Características faltantes: Supongamos que durante la clasificación, el valor de una de las características puede ser indeterminado, por ejemplo el ancho de un pez debido a una oclusión provocada por otro pez. ¿Cómo se podría compensar este problema?

* Se requiere de un entrenamiento Robusto*

Mapeo: Encontrar una transformación de las P medidas, típicamente, a un espacio de menor dimensión A : Conjunto de transformaciones

posibles, sea $\mathbf{x} = [x_1, x_2, \dots, x_p]$, entonces

$$J(A^*) = \max_{A \in \mathcal{A}} \{J(A)\},$$

se optimiza sobre las transformaciones posibles, y entonces se obtiene un nuevo vector

$$\mathbf{y} = A^*(\mathbf{x})$$

de menor dimensión (Reducción de dimensionalidad)

4.2. Razones para la extracción de características

Como ya se comentó anteriormente, una de las principales razones para llevar a cabo la extracción de características es reducir al máximo la cantidad de datos iniciales a procesar. En ese sentido, se estipulan las siguientes razones:

1: Proveer un conjunto relevante de características al clasificador, tratando de mejorar el desempeño de los clasificadores de estructura sencilla.

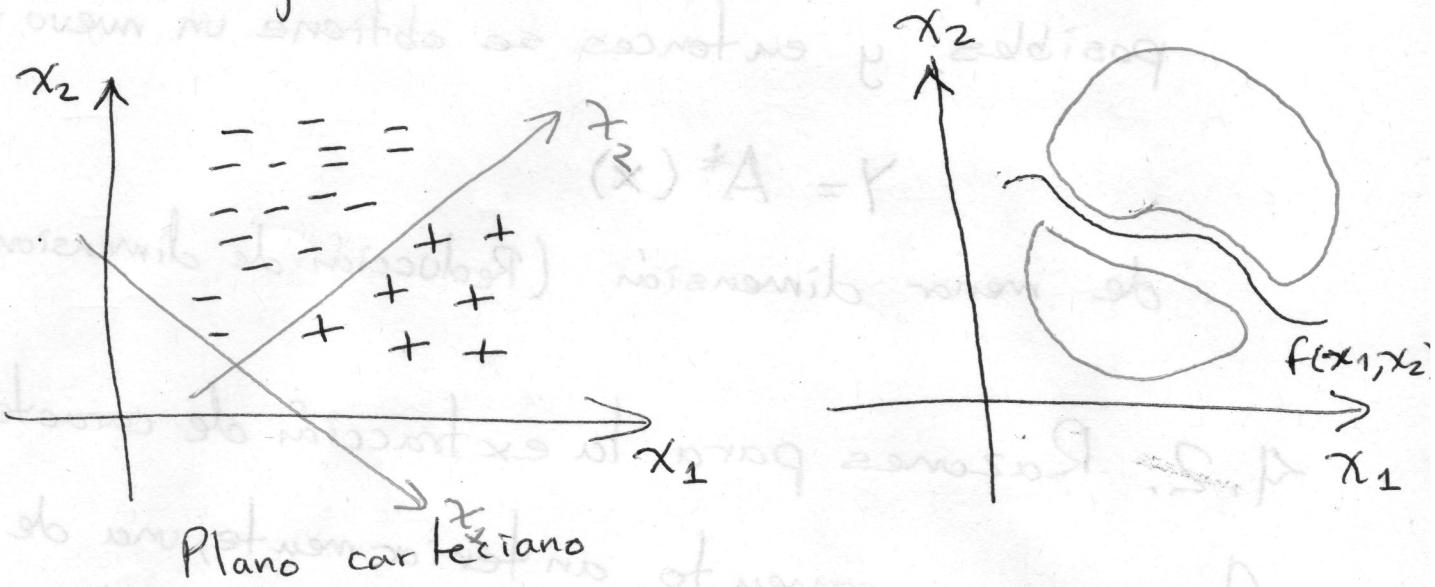
2: Reducir la redundancia en los datos.

3: Recuperar características latentes

significativas.

4º Generar o plasmar una mejor comprensión del proceso de generación de datos (Modelado)

5º Tener la capacidad de visualizar los conjuntos de datos



4.3. Algunas metodologías de extracción

Existe una gran cantidad de métodos de extracción de características, algunos muy arraigados al tipo de datos como en el caso de procesamiento de señales de voz, en donde tenemos al análisis por predicción lineal (LPC), o bien a los bancos de filtros para obtener coeficientes en escalas de Mel (MFCC).

Para el caso de procesamiento de imágenes, también se tienen otros tantos como los momentos de Hu, los coeficientes de Fourier,

ap A) Sabemos para cuales de los vectores
son los coeficientes del coseno discreto, etc. con

En años recientes, y de manera un poco más
general, se han popularizado los siguientes
métodos:

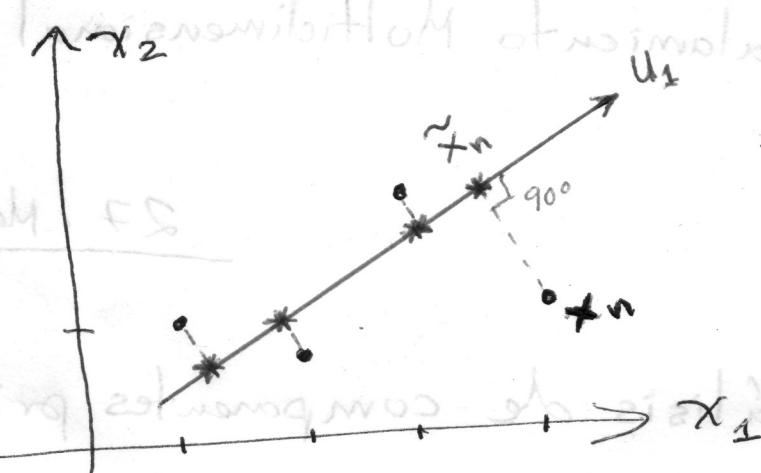
- Análisis de componentes principales (PCA),
- Análisis de discriminantes lineales (LDA),
- Análisis en componentes independientes (ICA),
- PCA mediante núcleos o Kernel PCA, y
- Escalamiento Multidimensional (MDS).

27 | Mayo | 2019

4.3.1. Análisis de componentes principales

El análisis de componentes principales (PCA por sus siglas en inglés - Principal Component Analysis), es una técnica ampliamente utilizada en aplicaciones que requieren una reducción de la dimensionalidad, compresión de datos con pérdidas, extracción de características y visualización de datos. También es conocida como la transformada de Karhunen-Loève (KLT).

Existen dos definiciones muy comunes de PCA que nos permiten llegar al mismo algoritmo. PCA puede ser definida como una proyección ortogonal de los datos sobre un espacio lineal de baja dimensión, conocido como el "sub-espacio principal", tal que la varianza de los datos proyectados se maximiza. De manera equivalente, también puede definirse como la proyección lineal que minimiza el costo promedio de proyección, tal que la distancia media cuadrática entre los datos y sus proyecciones es minimizada. El proceso de una proyección ortogonal se ilustra a continuación.



El sub-espacio principal corresponde a u_1 , de manera que los puntos negros proyectados en dicho subespacio maximizan la varianza de los puntos proyectados en azul.

También se puede minimizar la suma de los cuadrados de los errores de proyección (línea cortada).

- Formulación de minimización del error.
- El método PCA basado en la minimización del error de proyección, se discute a continuación, introduciremos un conjunto completo orthonormal de vectores base D -dimensionales $\{u_i\}$ en donde $i = 1, 2, \dots, D$, satisfaciendo la siguiente igualdad

oblig. los sistemas M tienen $U_i^T U_j = \delta_{ij}$.

Debido a que la base es completa, cada punto de los datos puede ser representado de manera exacta a partir de una combinación lineal de vectores base, esto es

$$x_n = \sum_{i=1}^M \alpha_{ni} U_i$$

en donde los coeficientes α_{ni} tomarán valores diferentes para diferentes puntos de datos. Esto corresponde simplemente a una rotación del sistema de coordenadas hacia un nuevo sistema definido por los vectores $\{U_i\}$, y las componentes originales D

$\{x_{n1}, x_{n2}, \dots, x_{nD}\}$ son repositionadas o sustituidas

por un conjunto equivalente $\{\alpha_{n1}, \alpha_{n2}, \dots, \alpha_{nD}\}$.

Tomando ahora el producto interno con U_j , y haciendo uso de la propiedad de ortogonalidad, se obtiene que $\alpha_{nj} = x_n^T U_j$, y por tanto podemos escribir

$$x_n = \sum_{i=1}^D (x_n^T U_i) U_i$$

Nuestra meta es sin embargo, aproximar estos puntos de datos utilizando una representación que involucra un número restringido tal que $M < D$ de variables que corresponden a una proyección en un sub-espacio de menor dimensión.

Entonces, el sub-espacio lineal M -dimensional puede ser representado utilizando los primeros M vectores base, y luego aproximamos cada punto de los datos \mathbf{x}_n , utilizando la siguiente ecuación

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^M z_{ni} \mathbf{u}_i + \sum_{i=M+1}^D b_i \mathbf{u}_i,$$

en donde los $\{z_{ni}\}$ dependen de los puntos de datos particulares, mientras que los $\{b_i\}$ son constantes que toman el mismo valor para todos los puntos de datos.

Tenemos libertad de elegir los $\{\mathbf{u}_i\}$, los $\{z_{ni}\}$, y los $\{b_i\}$ de tal modo que se puede minimizar la

distorsión introducida debido a la reducción en dimensionalidad. Como nuestra medida de distorsión

utiliza una distancia cuadrática entre los puntos de

datos originales \mathbf{x}_n y su aproximación $\tilde{\mathbf{x}}_n$, y el

promedio sobre el conjunto de datos, entonces nuestra meta es minimizar el siguiente criterio

$$J = \frac{1}{N} \sum_{n=1}^N \|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|^2.$$

Primeramente se considera la minimización con respecto de

$\{z_{ni}\}$, sustituyendo para $\tilde{\mathbf{x}}_n$, luego derivando el

criterio con respecto de z_{ni} , e igualando con cero.

Haciendo uso también de las relaciones de ortonormalidad, tenemos que

$$z_{ni} = \mathbf{x}_n^T \mathbf{u}_i,$$

en donde $j=1, 2, \dots, M$: De manera similar, a hora derivando el criterio J con respecto de $\{b_i\}$ e igualando con cero, y utilizando las condiciones de ortonormalidad, entonces

$$b_j = \tilde{x}^T u_j,$$

$$T(\tilde{x} - x_n)(\tilde{x} - x_n)^T = \sum_{i=1}^N \tilde{x}^T u_i u_i^T = 2, \quad \tilde{x} = \frac{1}{N} \sum_{n=1}^N x_n,$$

en donde $j = M+1, M+2, \dots, D$. Si sustituimos para x_n y b_i , y hacemos uso de la expansión para x_n , obtenemos

$$x_n - \tilde{x}_n = \sum_{i=M+1}^D \{(x_n - \tilde{x})^T u_i\} u_i,$$

de donde podemos ver que el vector de desplazamiento de x_n a \tilde{x}_n cae en el espacio ortogonal del sub-espacio principal, dado que es una combinación lineal de $\{u_i\}$ para $i = M+1, M+2, \dots, D$. Esto era de esperarse debido a que los puntos proyectados \tilde{x}_n deben de caer en el sub-espacio principal, y los podemos mover de ahí libremente dentro del sub-espacio, y el error mínimo estará dado por la proyección ortogonal.

De manera general, se puede obtener una expresión del criterio para la medida de distorsión J en términos de $\{u_i\}$ solamente, esto es

$$J = \frac{1}{N} \sum_{n=1}^N \sum_{i=M+1}^D (x_n^T u_i - \bar{x}^T u_i)^2$$

$$= \sum_{i=M+1}^D u_i^T S u_i,$$

$$\text{en donde } S = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})(x_n - \bar{x})^T,$$

que es la matriz de covarianza de los datos.

La tarea sigue siendo minimizar J con respecto de $\{u_i\}$, que corresponde a una minimización condicionada. Si consideramos $D=2$, y $M=1$

para datos en 2 dimensiones, entonces el sub-espacio principal será para $M=1$. Si elegimos la dirección u_2 de manera que debemos minimizar $J = u_2^T S u_2$

con una condición de normalización $u_2^T u_2 = 1$. Y utilizando un "multiplicador de Lagrange" λ_2 para reforzar el condicionamiento, entonces se considera la minimización de

$$\tilde{J} = u_2^T S u_2 + \lambda_2 (1 - u_2^T u_2).$$

y derivando el criterio con respecto de u_2 e igualando con cero, tenemos que

$$S u_2 = \lambda_2 u_2,$$

de modo general, la solución se puede escribir

$$S u_i = \lambda_i u_i$$

para $i=1, 2, \dots, D$, en donde los $\{u_i\}$ serán denominados eigenvectores y se elijan como bases orto normales. Y el valor correspondiente de la medida de distorsión estará dado por

$$J = \sum_{i=M+1}^D \lambda_i,$$

que corresponde simplemente a la suma de los eigenvalores de dichos eigenvectores los cuales son ortogonales al sub-espacio principal.

(XI)

(A1)

4.3.2. Kernel PCA.

La técnica de sustitución de núcleos o **Kernels** nos permite construir un algoritmo expresado en términos de productos escalares de la forma $x^T x'$ y generalizar dicho algoritmo reemplazando los productos escalares con un Kernel no lineal. Se aplica la técnica de sustitución de Kernel dentro del análisis de componentes principales, obteniendo así una generalización no lineal que se denomina **Kernel PCA**.

Consideremos nuevamente un conjunto de datos de $\{x_n\}$ observaciones, donde $n=1, 2, \dots, N$, en un espacio de dimensionalidad D . Ahora, asumiremos que ya se ha obtenido la media de la muestra para cada uno de los vectores $\{x_n\}$, de modo que $\sum_n x_n = 0$. El primer paso es expresar el PCA convencional de tal manera que los vectores $\{x_n\}$ aparezcan

A3

restarle 100

sólo en forma de productos escalares $x_n^T x_m$.

Recordemos que las componentes principales están definidas por los vectores (eigenvectores) u_i de la matriz de covarianza S de la muestra, en donde $i = 1, 2, \dots, D$. También, la matriz de covarianza de la muestra de dimensión $D \times D$, está definida por

$$S = \frac{1}{N} \sum_{n=1}^N x_n^T x_n$$

y los eigenvectores están normalizados de tal manera que

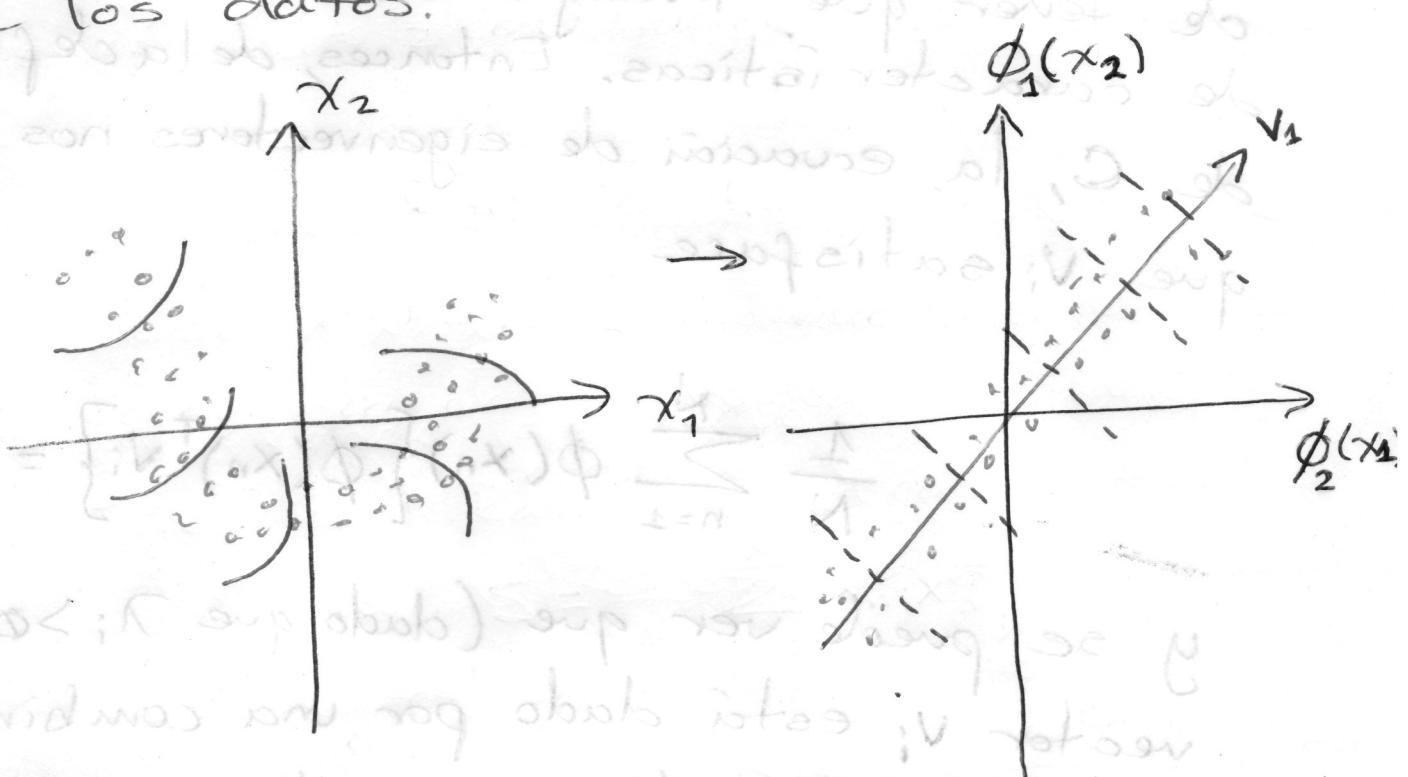
$$u_i^T u_i = 1.$$

Ahora, consideremos la transformación no lineal $\phi(x)$ hacia un espacio de características M -dimensional, entonces cada punto x_n de los datos se proyecta sobre un punto $\phi(x_n)$. Ahora, podemos aplicar el análisis estándar PCA en este nuevo espacio de características, lo cual define

A3

PA

implícitamente el modelo de componentes principales no lineal en el espacio original de los datos.



Por el momento, asumamos que el conjunto de datos proyectados también tiene media cero, así que $\sum_n \phi(x_n) = 0$. La matriz de covarianza de dimensión $M \times M$, C en el nuevo espacio de características está dada por

$$C = \frac{1}{N} \sum_{n=1}^N \phi(x_n) \phi(x_n)^T$$

y esta tiene su expansión en eigenvectores

definida por

$$Cv_i = \lambda_i v_i$$

(A4)

EA

con $i = 1, 2, \dots, M$. Nuestra meta es resolver este problema de eigenvalores sin necesidad de tener que trabajar en el nuevo espacio de carácterísticas. Entonces, de la definición de C , la ecuación de eigenvectores nos dice que v_i satisface

$$\frac{1}{N} \sum_{n=1}^N \phi(x_n) \{ \phi(x_n)^T v_i \} = \lambda_i v_i,$$

y se puede ver que (dado que $\lambda_i > 0$) el vector v_i está dado por una combinación

lineal de los $\phi(x_n)$ y entonces puede escribirse de la forma

$$v_i = \sum_{n=1}^N a_{in} \phi(x_n).$$

Sustituyendo esta expansión en la ecuación de eigenvectores anterior, obtenemos

$$\frac{1}{N} \sum_{n=1}^N \phi(x_n) \phi(x_n)^T \sum_{m=1}^N a_{im} \phi(x_m) = \lambda_i \sum_{m=1}^N a_{im} \phi(x_m),$$

Y ahora, la clave es expresar esta ecuación en términos de una función $K(x_n, x_m) = \phi(x_n)^T \phi(x_m)$, si se multiplica

05) Jun 7 2019
A5 en ambos lados de la ecuación por $\phi(x_i)^T$

$$\frac{1}{N} \sum_{n=1}^N k(x_i, x_n) \sum_{m=1}^N a_m k(x_n, x_m) = \lambda_i \sum_{n=1}^N a_n k(x_i, x_n)$$

ya que $k(x_n, x_m) = \phi(x_n)^T \phi(x_m)$, y escribiendo

la ecuación en forma matricial, tenemos que

$$K^2 a_i = \lambda_i N K a_i,$$

en donde a_i es un vector N -dimensional columna con elementos a_{ni} para $n=1, 2, \dots, N$.

Entonces, podemos encontrar soluciones para a_i brindando solución al siguiente problema de eigenvalores

$$K a_i = \lambda_i N a_i,$$

en donde hemos quitado o eliminado un factor

A) $\times K$. Nótese que las soluciones de las dos ecuaciones anteriores sólo difieren por los eigenvectores de K que tienen eigenvalores iguales a cero, por lo que no se afecta la proyección de las componentes principales.

La condición de normalización para los coeficientes a_i se obtiene estableciendo que los eigenvectores en el espacio de características sean normalizados, esto es

$$1 = \mathbf{v}_i^T \mathbf{v}_i = \sum_{n=1}^N \sum_{m=1}^N a_{in} a_{im} \phi(x_n)^T \phi(x_m)$$

$$= a_i^T K a_i = \lambda_i N a_i^T a_i.$$

Habiendo solucionado este problema de normalización de eigenvectores, las proyecciones resultantes de la componente principal también se pueden expresar en términos de una función Kernel de manera que, la proyección de un punto x sobre el i -ésimo eigenvector estará dada por

$$y_i(x) = \phi(x)^T \mathbf{v}_i = \sum_{n=1}^N a_{in} \phi(x)^T \phi(x_n)$$

$$= \sum_{n=1}^N a_{in} K(x, x_n),$$

y entonces $y_i(x)$ se expresa en términos de K .

(B)

(A)

En el espacio original D -dimensional de x existen D eigenvectores ortogonales y entonces podemos encontrar las D componentes principales lineales. La dimensionalidad M del espacio de características, sin embargo, podría ser mucho mayor a D (inclusiva infinita), y entonces sería posible de determinar un número ~~de~~ de componentes principales no lineales que podría exceder a D .

Notese sin embargo, que el número de eigenvalores no cero no puede exceder a N , que corresponde a la cantidad de datos, aún y cuando $M > N$, esto se debe a que la matriz de covarianzas en el espacio de características tiene un rango a lo más igual a N . Esto se refleja en el hecho de que Kernel PCA involucra una expansión de vectores eigen con matriz K de $N \times N$.

Anteriormente se asumió que el conjunto de datos proyectados por $\phi(x_n)$ tiene media igual a cero, en general eso podría no ser el caso. No podemos simplemente calcularla y eliminarla, dado que deseamos evitar trabajar directamente en el espacio X , y a cambio queremos calcular en el espacio de características, y nuevamente aquí debemos formular un algoritmo en términos puramente de una función Kernel. Los puntos de los datos proyectados después de un centrado, se pueden escribir como $\tilde{\phi}(x_n)$, y estarán dados por

$$\tilde{\phi}(x_n) = \phi(x_n) - \frac{1}{N} \sum_{i=1}^N \phi(x_i),$$

y los elementos correspondientes de la matriz Gram estarán dados por

$$\tilde{K}_{nm} = \tilde{\phi}(x_n)^T \tilde{\phi}(x_m)$$

$$= \phi(x_n)^T \phi(x_m) - \frac{1}{N} \sum_{l=1}^N \phi(x_n)^T \phi(x_l)$$

$$- \frac{1}{N} \sum_{l=1}^N \phi(x_l)^T \phi(x_m) + \frac{1}{N^2} \sum_{j=1}^N \sum_{l=1}^N \phi(x_j)^T \phi(x_l)$$

$$\tilde{K}_{nm} = k(x_n, x_m) - \frac{1}{N} \sum_{l=1}^N k(x_l, x_m)$$

$$- \frac{1}{N} \sum_{l=1}^N k(x_n, x_l) + \frac{1}{N^2} \sum_{j=1}^N \sum_{l=1}^N k(x_j, x_l)$$

D) Y re-escribiéndola en forma matricial, tendremos que

$$\tilde{K} = K - \mathbf{1}_N K - K \mathbf{1}_N + \mathbf{1}_N K \mathbf{1}_N,$$

en donde $\mathbf{1}_N$ denota una matriz de $N \times N$

elementos en donde cada elemento toma el valor $1/N$.

Luego, podemos evaluar \tilde{K} utilizando sólo la función Kernel y entonces utilizamos \tilde{K} para determinar los eigenvectores y eigenvalores.

Notese que el algoritmo estandar de PCA

corresponde a un caso especial de Kernel PCA cuando se utiliza un kernel lineal $K(x, x') = x^T x'$.

Ejemplo: Si se aplica kernel PCA a un conjunto de datos sintéticos de Schölkopf (1998), y si se utiliza un kernel Gaussiano de la forma

$$K(x, x') = \exp(-\|x - x'\|^2 / 0.1).$$

Se obtienen líneas de contorno que corresponden a las proyecciones en componentes principales, definidas como

$$\phi(x)^T v_i = \sum_{n=1}^N a_{in} k(x, x_n),$$

y las cuales son constantes.

(E) En este caso, se puede ver una desventaja de Kernel PCA y ésta es que determinar los eigenvectores de $N \times N$ para la matriz \tilde{K} en lugar de los de la matriz S de $D \times D$, hace que se tengan costos de cálculo mayores pues se tendrán conjuntos de datos muy grandes.

También, se hace notar que para el caso de PCA (estándar), la idea es retener un número reducido de $L < D$ de eigenvectores y luego aproximar el vector de datos x_n por sus proyecciones \hat{x}_n sobre el espacio principal

2-dimensional, de modo que \mathbf{z} es:

$$\hat{\mathbf{z}}_n = \sum_{i=1}^L (\mathbf{x}_n^T \mathbf{u}_i) \mathbf{u}_i$$

• (Foto) en el caso de Kernel PCA, en general no será posible de realizar.

4.3.3. Análisis en componentes independientes

En este caso, comenzamos considerando modelos

⑦ en los cuales las variables observadas están relacionadas de manera lineal con las variables latentes, pero las distribuciones de las variables latentes son no-Gaussianas. A este tipo de modelos, se les conoce como análisis en componentes independientes (ICA), y provienen de la siguiente consideración de la distribución de las variables latentes que factorizan de modo que

$$P(\mathbf{z}) = \prod_{j=1}^M p(z_j)$$

Al ICA busca las direcciones más independientes en lugar de minimizar los errores de modelado como en el caso de PCA. Estas ideas se utilizan en procesamiento de señales en la separación de fuentes, como por ejemplo, en la separación de mezclas de voz, o en la separación de canales de señales electroencefalográficas (EEG).

El método se basa en la independencia estadística de dichas fuentes, y las hipótesis de que las componentes son ortogonales no se toma en cuenta para la separación.

⑤ Sea $\mathbf{x} = \{x_1, x_2, \dots, x_m\}$ un vector de m componentes que corresponde a los datos de entrenamiento, asumimos que este vector se produce como una combinación lineal (mezcla) de m señales independientes que denotaremos como el vector \mathbf{s} , $\mathbf{s} = \{s_1, s_2, \dots, s_m\}$, es decir, tenemos el siguiente sistema de ecuaciones

$$x_1 = a_{11}s_1 + a_{12}s_2 + \dots + a_{1m}s_m,$$

$$x_2 = a_{21}s_1 + a_{22}s_2 + \dots + a_{2m}s_m,$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots$$

$$x_m = a_{m1}s_1 + a_{m2}s_2 + \dots + a_{mm}s_m.$$

Los coeficientes a_{ij} corresponden a una matriz A , que se conoce como "matriz de la mezcla", al vector x lo llamaremos el vector de mezclas y s es el vector de componentes independientes.

Entonces, podemos escribir el sistema anterior de forma matricial como

$$x = As,$$

en donde en la práctica sólo conocemos x que se genera al muestrear alguna señal o evento.

④ Luego, el método ICA consiste en aplicar un algoritmo que nos permita encontrar una matriz de mezcla W de tal modo que

$$y = Wx,$$

nos brinde una buena aproximación al vector s , esto es $y \approx s$, de manera que $|y - s| \rightarrow 0$.

estadísticas para análisis de series espacio-temporales)

propiedades hidráulicas y porque es importante analizarla. (Menzione algunas pruebas

importancia en el riego y drenaje agrícola y que entiende por variabilidad espacial de las

8. Explique qué es la caracterización hidrodinámica de suelos y porque es de fundamental

Las condiciones para un buen funcionamiento del método ICA son las siguientes:

1- Independencia estadística: La independencia estadística es la idea primordial para todos los usos fructíferos del método, queremos encontrar dentro de nuestras señales mezcladas, aquellas que son estadísticamente más independientes con respecto a las demás. Podemos definir la independencia estadística como sigue:

(I) Sean x_1, x_2, \dots, x_m un conjunto de v.a. con función de densidad de probabilidad $p(x_1, x_2, \dots, x_m)$, entonces estas variables son mutuamente independientes si

$$P(x_1, x_2, \dots, x_m) = p_1(x_1) p_2(x_2) \dots p_m(x_m),$$

en donde $p_i(x_i)$ es la función de densidad marginal de x_i . Se asumirá para la construcción del método que cada una de las señales si sea estadísticamente independiente con respecto a las demás señales.

2.- La matriz A debe ser cuadrada: Para la construcción del método es necesario que la matriz de mezcla A sea cuadrada y de rango completo, es decir, que el número de fuentes si sea igual al número de mezclas x :

3.- Se asume que el experimento está libre de ruido: Todas las señales que se muestran de las señales mezcladas X deben tener solamente combinaciones lineales de las señales independientes S , no debe haber información proveniente de ruido externo.

④ 4.- Los datos deben estar centrados: su medida prom. debe ser cero. Aunque en general, esta condición es difícil de obtener, y se puede hacer un preprocessado de los datos para re-centrarlos, a dicho proceso se le denomina "blanqueo".

5.- La función de densidad de probabilidad es no Gaussiana: Las señales fuente independientes deben tener función de densidad de probabilidad no Gaussiana, de esta forma se garantiza que las componentes independientes efectivamente pueden ser separadas.

El primer paso para la construcción del método es entonces "blanquear" los datos, es decir centrar los datos del vector de señales x , para esto se resta la media de cada una de las componentes observadas. Después de esto, a partir de la matriz de covarianza de los datos x , se elimina la correlación entre cada una de las señales observadas.

(K)