Tarea 3 Reconocimiento de Patrones

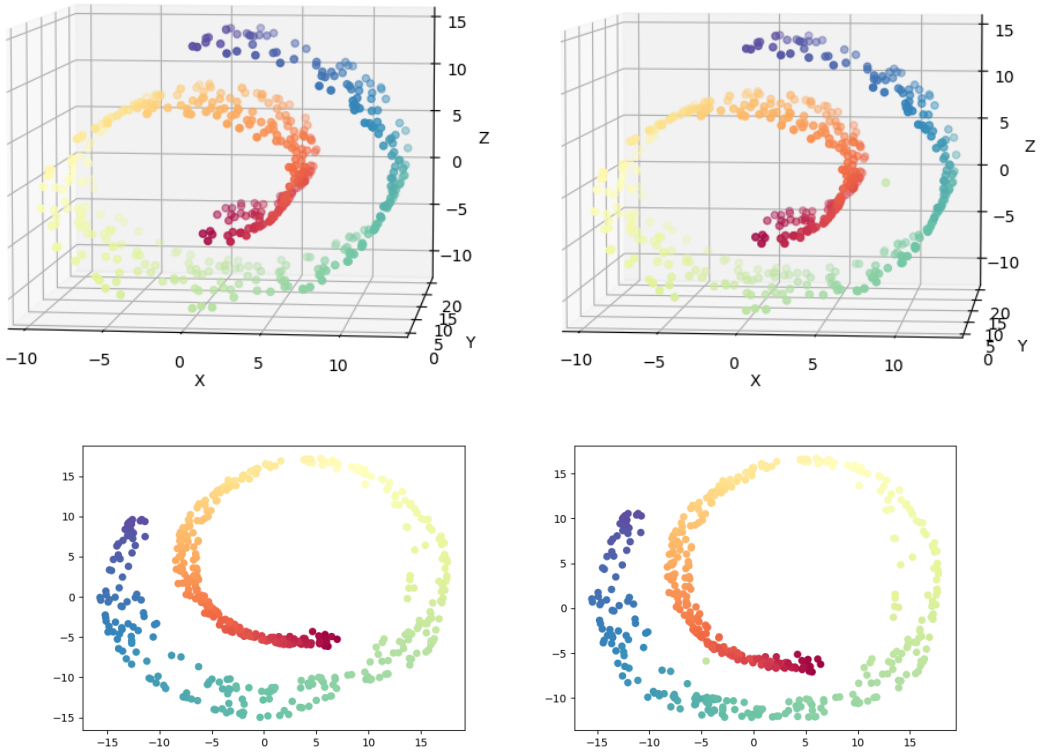
Yamil Ernesto Morfa Avalos

27 de febrero de 2021

B. Preguntas cortas

1. En el ejemplo visto para ilustrar ISOMAP definimos que dos observaciones x_i y x_j están conectadas s.i.i x_i está entre los k-vecinos más cercanos de x_j y a su vez x_j está entre los k-vecinos más cercanos de x_i. La idea es mostrar que la adición de una sola observación en este ejemplo puede destruir por completo el desenrollamiento.

Figura 1:



En la figura 1 se muestra este fenómeno. Se puede observar que la estructura local de la vecindad donde se proyecta el punto añadido si ha cambiado. Esto sucede por que para el valor seleccionado de $k = 10$, el punto añadido cae entre los vecinos más cercanos de valores con los cuales no está relacionado.

2. El teorema de Rao plantea que si $\mathbb{F}$ es una matriz simétrica de rango d con SVD

$$\mathbb{F} = \sum_{i=1}^{d} \lambda_i v_i v_i^T$$

entonces

$$\mathbb{G} = \arg \min_{rang(B)=p} \|\mathbb{F} - B\|_F = \sum_{i=1}^{p} \lambda_i v_i v_i^T \quad con\ p < d$$

Veamos que el error $\|\mathbb{F} - \mathbb{G}\|_F^2 = \sum_{i=p+1}^{d} \lambda_i^2$

$$\|\mathbb{F} - B\|_F^2 = trz\left\{(\mathbb{F} - \mathbb{G})^T (\mathbb{F} - \mathbb{G})\right\} = trz\left\{\mathbb{F}^T\mathbb{F} - \mathbb{G}^T\mathbb{F} - \mathbb{F}^T B + \mathbb{G}^T\mathbb{G}\right\} =$$

$$= trz\left\{\left(\sum_{i=1}^{d} \lambda_i v_i v_i^T\right)^T \left(\sum_{i=1}^{d} \lambda_i v_i v_i^T\right) - \left(\sum_{i=1}^{p} \lambda_i v_i v_i^T\right)^T \left(\sum_{i=1}^{d} \lambda_i v_i v_i^T\right) - \left(\sum_{i=1}^{d} \lambda_i v_i v_i^T\right)^T \left(\sum_{i=1}^{p} \lambda_i v_i v_i^T\right) + \left(\sum_{i=1}^{p} \lambda_i v_i v_i^T\right)^T \left(\sum_{i=1}^{p} \lambda_i v_i v_i^T\right)\right\} =$$

dado que tanto $\mathbb{F}$ y $\mathbb{G}$ son simétricas:

$$= trz\left\{\sum_{j=1}^{d}\left[\lambda_j v_j v_j^T \left(\sum_{i=1}^{d} \lambda_i v_i v_i^T\right)\right] - \left[\sum_{j=1}^{p} \lambda_j v_j v_j^T \left(\sum_{i=1}^{d} \lambda_i v_i v_i^T\right)\right] - \left[\sum_{j=1}^{d} \lambda_j v_j v_j^T \left(\sum_{i=1}^{p} \lambda_i v_i v_i^T\right)\right] + \left[\sum_{j=1}^{p} \lambda_j v_j v_j^T \left(\sum_{i=1}^{p} \lambda_i v_i v_i^T\right)\right]\right\} =$$

$$= trz\left\{\sum_{j=1}^{d}\sum_{i=1}^{d} \lambda_i\lambda_j v_j v_j^T v_i v_i^T - \sum_{j=1}^{p}\sum_{i=1}^{d} \lambda_i\lambda_j v_j v_j^T v_i v_i^T - \sum_{j=1}^{d}\sum_{i=1}^{p} \lambda_i\lambda_j v_j v_j^T v_i v_i^T + \sum_{j=1}^{p}\sum_{i=1}^{p} \lambda_i\lambda_j v_j v_j^T v_i v_i^T\right\} =$$

$$= trz\left\{\sum_{i=1}^{d} \lambda_i^2 v_i v_i^T - \sum_{i=1}^{p} \lambda_i^2 v_i v_i^T - \left(\sum_{i=1}^{p} \lambda_i^2 v_j v_i^T + \cancelto{0}{\sum_{j=p+1}^{d}\sum_{i=1}^{p} \lambda_i\lambda_j v_j v_j^T v_i v_i^T}\right) + \sum_{i=1}^{p} \lambda_i^2 v_i v_i^T\right\} =$$

$$= trz\left\{\sum_{i=1}^{d} \lambda_i^2 v_i v_i^T - \sum_{i=1}^{p} \lambda_i^2 v_i v_i^T\right\} = trz\left\{\sum_{i=p+1}^{d} \lambda_i^2 v_i v_i^T\right\} = \sum_{i=p+1}^{d} \lambda_i^2$$

dado que la traza de una matriz es la suma de sus valores propios.

3. Supongamos que P^1 y P^2 se distribuyen Bernullis con parámetros θ_1 y θ_2 respectivamente. Tenemos que para distribuciones discretas la distancia de Kullback-Leibler es:

$$d\left(P^1, P^2\right) = \sum_i P_i^1 \log \frac{P_i^1}{P_i^2}$$

esto es:

$$d\left(P^1, P^2\right) = \sum_i \theta_1^i \left(1-\theta_1\right)^{1-i} \log \frac{\theta_1^i \left(1-\theta_1\right)^{1-i}}{\theta_2^i \left(1-\theta_2\right)^{1-i}} =$$

$$= \sum_i \theta_1^i \left(1-\theta_1\right)^{1-i} \left\{i \log \frac{\theta_1}{\theta_2} + (1-i) \log \frac{(1-\theta_1)}{(1-\theta_2)}\right\} =$$

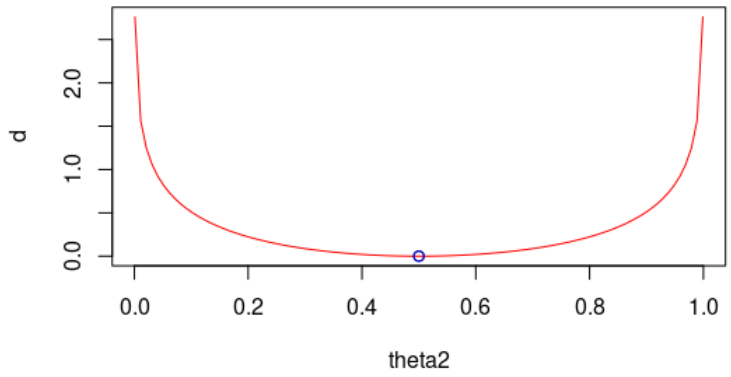
$$= \log \frac{\theta_1}{\theta_2} \left\{\sum_i i\theta_1^i \left(1-\theta_1\right)^{1-i}\right\} + \log \frac{(1-\theta_1)}{(1-\theta_2)} \left\{\sum_i (1-i)\, \theta_1^i \left(1-\theta_1\right)^{1-i}\right\} =$$

$$= \log \frac{\theta_1}{\theta_2} \left\{ \sum_i i P_i^1 \right\} + \log \frac{(1-\theta_1)}{(1-\theta_2)} \left\{ \sum_i (1-i)\, \theta_1^i\, (1-\theta_1)^{1-i} \right\} = *$$

ahora dado que $\sum_i i P_i^1 = E\left[P^1\right] = \theta_1$ y $\sum_i (1-i)\, \theta_1^i\, (1-\theta_1)^{1-i} = E\left[Ber\,(1-\theta_1)\right] = 1-\theta_1$

$$d\left(P^1, P^2\right) = \theta_1 \log \frac{\theta_1}{\theta_2} + (1-\theta_1) \log \frac{(1-\theta_1)}{(1-\theta_2)}$$

Figura 2:



En la figura 2 se muestra, para $\theta_1 = 0{,}5$, la distancia $d\left(P^1, P^2\right)$ en función de θ_2. Veamos que a medida que θ_2 se acerca a 0,5 las Bernullis se parecen más, esto se puede observar en la figura dado que para valores cercanos a 0,5 se puede ver como $d\left(P^1, P^2\right)$ se acerca a 0. Por tanto podemos interpretar que, en efecto, $d\left(P^1, P^2\right)$ mide la distancia entre las Bernullis.

4. Sea S un conjunto finito. Definimos como medida de similitud entre dos subconjuntos A y B de S:

$$K(A, B) := \#\left(A \bigcap B\right)$$

Definamos $\Phi : \mathcal{P}(S) \to \mathbb{R}^n$ donde $\mathcal{P}(S)$ es el conjunto potencia de S y $n = \#S$, como:

$$\Phi_{i \in I}(A) = 1\,(s_i \in A) = \begin{cases} 1 & si\ s_i \in A \\ 0 & si\ s_i \notin A \end{cases}$$

donde I es un conjunto de subíndices de cardinalidad n con el cual indexamos los elementos de S, s_i. Notemos que para dos subconjuntos A, B tenemos que $\Phi(A) \in \mathbb{R}^n$ es un vector con 1 en las posiciones donde "hay elementos" y 0 en el resto y lo mismo para $\Phi(B)$. O sea que

$$\langle \Phi(A), \Phi(B) \rangle = \sum_{i \in I} \Phi_i(A)\, \Phi_i(B) = k < n$$

donde k es el numero de veces en que $\Phi_i(A)\Phi_i(B) = 1 \Longrightarrow \Phi_i(A) = 1 \wedge \Phi_i(B) = 1$ o equivalente:

$$s_i \in A \quad \wedge \quad s_i \in B \Longrightarrow s_i \in A \bigcap B$$

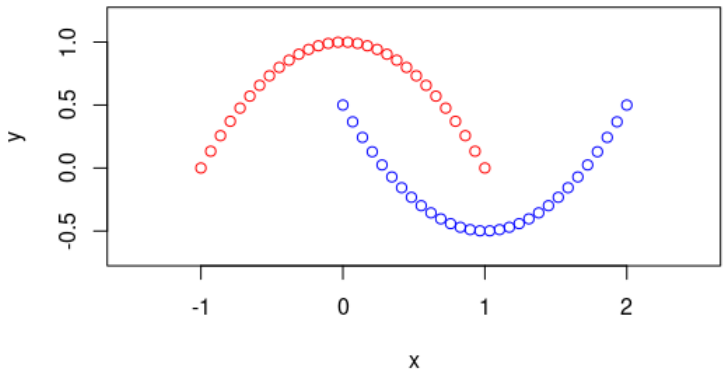
Luego

$$K(A,B) = \#\left(A \bigcap B\right) = \langle \Phi(A), \Phi(B) \rangle$$

C. Análisis de datos

1. Veamos un ejemplo de $KPCA$ con un kernel centrado de base radial con parámetro σ, $\mathbb{K}_c = \mathbb{C}\mathbb{K}\mathbb{C}$ con $\mathbb{K}_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma}\right)$ y $\mathbb{C}$ la matriz para centrar los datos. En primer lugar creemos los dato par este ejemplo.

Figura 3:

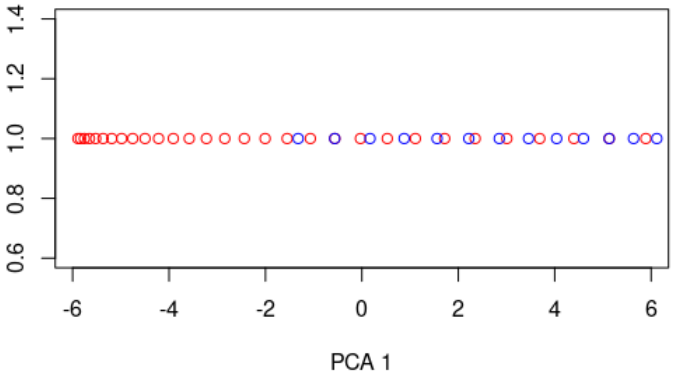


En la figura 3 se muestran los datos generados mediante dos parábolas:

$$x_{i2} = -x_{i1}^2 + 1 \qquad x_{i1} \in [-1,\ 1]\ \ i = 1, 2, ..., 30\ equiespaciados$$
$$x_{j2} = (x_{j1} - 1)^2 - 0{,}5 \qquad x_{j1} \in [0,\ 2]\ \ j = 1, 2, ..., 30\ equiespaciados$$

Si aplicamos PCA para estos datos y proyectamos sobre la primera componente, vemos que:

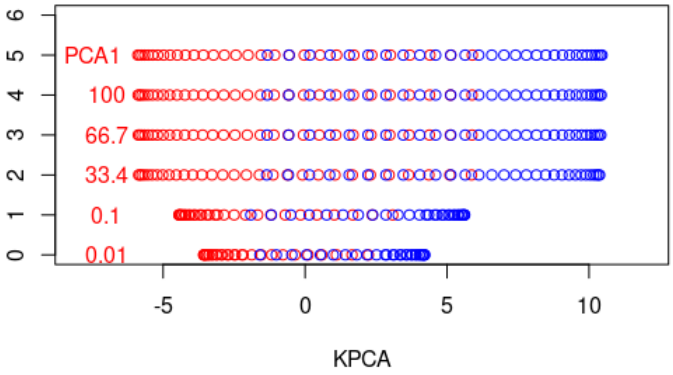
Figura 4:



En la figura 4 se muestra a modo de ilustración la proyección de los datos sobre la primera componente principal. Esta se ha graficado sobre la recta $y = 1$ para mayor visualización.

Para realizar el $KPCA$ se ha construido la función kernel como se define anteriormente y luego se ha obtenido matriz $\mathbb{K}_c$. Se le ha realizado la descomposición SVD para obtener la primera componente de $KPCA$.

Figura 5:



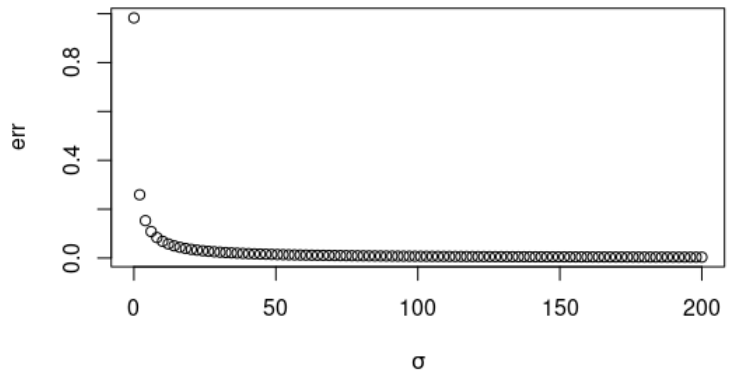
En la figura 5 se muestra una comparación de los datos proyectados sobre las componentes obtenidas para distintos valores de σ (los valores que aparecen en rojo). Para una mejor visualización se han graficado sobre las rectas $y = 0$, $y = 1$, $y = 2$, $y = 3$, $y = 4$, $y = 5$ las proyecciones de los datos para los valores de $\sigma = 0{,}01$, $\sigma = 0{,}1$, $\sigma = 33{,}4$, $\sigma = 66{,}7$, $\sigma = 100$ y $PCA1$ usual respectivamente. Es fácil notar que a medida que σ aumenta la solución de $KPCA$

tiende a ser la solución usual de PCA. Esto se comprueba en la figura 6 donde representamos

$$err = \|pca1 - kpca1\,(\sigma)\|_2$$

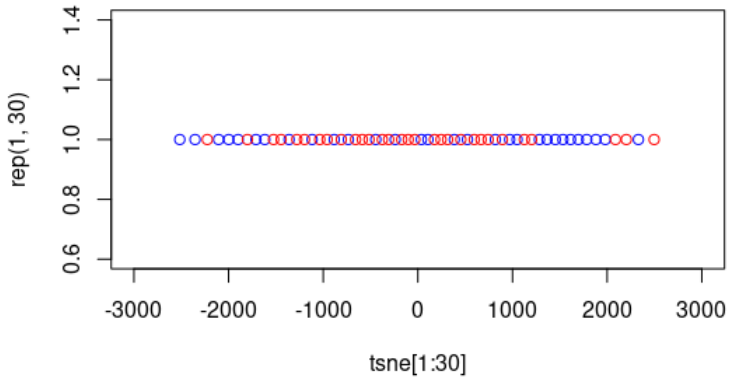
para un conjunto mas amplio de valores de σ

Figura 6:



Para esto se tomaron 100 valores de σ equiespaciados de 0 hasta 200. El valor mínimo obtenido para esta elección del error es $3{,}7e^{-3}$ y este sigue decreciendo hasta 0 a medida que $\sigma \to \infty$.

Figura 7:

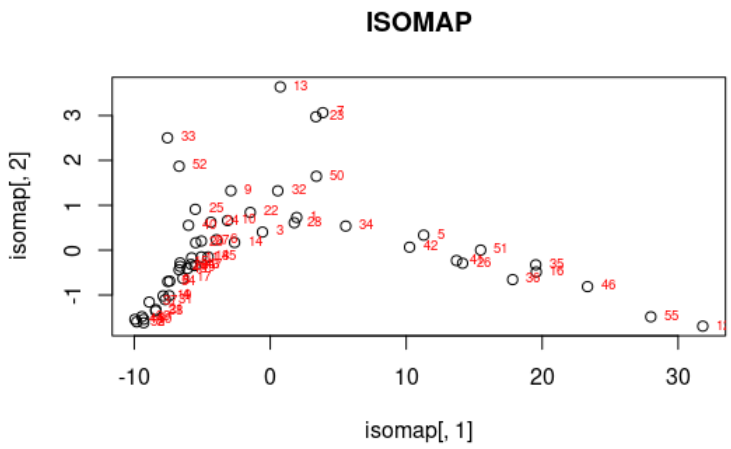


En la figura 7 se muestra la proyección para $T - SNE$. Podemos ver que la solución no es buena. No realiza una separación adecuada de los datos. Esto se debe a que que $T - SNE$ es muy sensible al parámetro perplexity y al

punto de arranque

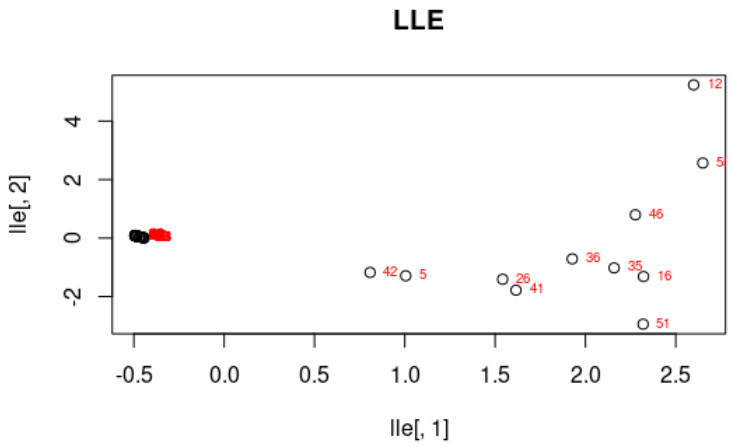
2. Trataremos de encontrar visualizaciones ilustrativas para los datos usados en la primera tarea *deport.dat*. Esto es un conjunto de $n = 55$ observaciones de los tiempos en segundo de $p = 8$ eventos de atletismo. Se usarán los métodos $ISOMAP$, LLE, $T - SNE$ y SOM

Figura 8:



En la Fig 8 se muestra la proyección sobre un espacio de dimensión 2 de $ISOMAP$ para $k = 5$ vecinos.

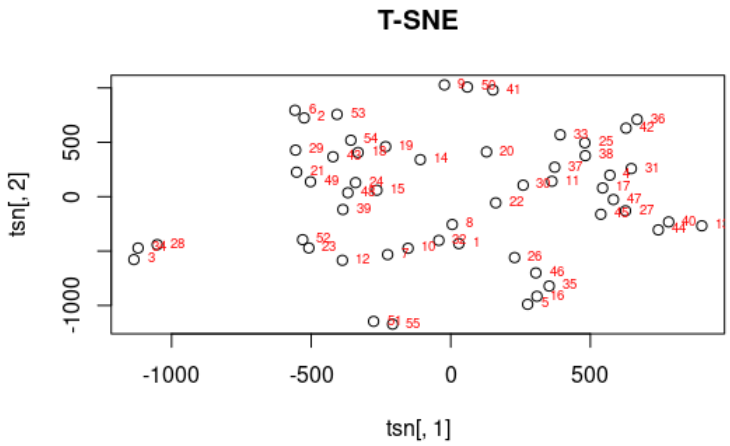
Figura 9:



En la fig 9 se muestra la proyección sobre un espacio de dimensión 2 para LLE para $k = 5$ vecinos. Podemos ver que la mayor cantidad de datos están concentrados en una vecindad pequeña del espacio. Esto se puede interpretar que los datos que allí se encuentran son de alguna forma muy similares. Se pueden observar grupos aislados como

son {*bermuda*, *philippi*}; {*indonesi*, *png*}; {*domrep*, *malaysia*, *mauritiu*}; {*singapor*}; {*thailand*}; {*wsamoa*}; {*cookis*} siendo este ultimo el más aislado (atípico) de todos. Entre los elementos de estos grupos se podría decir que hay cierta similaridad pero de manera general todos los grupos están "lejos" de la mayoría de los datos

Figura 10:



En la fig 10 se muestra la proyección sobre un espacio de dimensión 2 para $T-SNE$ para un valor de $perplexity = 5$. En este caso vemos que los datos no siguen cierta estructura como en los casos anteriores. Esto se puede deber a que $T - SNE$ es muy inestable y sensible al parámetro de perplexity.