Tarea 02 - Reconocimiento de patrones

Giovanni Gamaliel López Padilla

Problema 2.1

Sea $\{xi\}$ un conjunto de n vectores d dimensional. Definimos la matriz Kernel $[K_{i,j}]$ con $K_{i,j} = \langle x_i, x_j \rangle$ y $\mathbb{D}^2$ la matriz de distancias al cuadrada correspondiente. Verifica la identidad que usamos en clase:

$$\mathbb{D}^2 = c1^t + 1c^t - 2\mathbb{X}\mathbb{X}^t$$

con 1 un vector de unos de longitud n y c el vector de longitud n con elementos $(\mathbb{K}_{i,i})_{i=1}^n$

Sea X una matriz con elementos x_{ij}, entonces, la matriz XX^T se puede escribir de la siguiente manera:

$$(XX^T)_{ij} = \sum_{k=1}^{d} x_{ik}x_{jk}$$

Con esto, el producto $1c^T$, se puede calcular como:

$$(1^Tc)_{ij} = \sum_{k=1}^{d} x_{jk}^2$$

De igual manera, el producto c^T1, se puede calcular como:

$$(c^T1)_{ij} = \sum_{k=1}^{d} x_{ik}^2$$

entonces el elemento ij de la matriz $\mathbb{D}^2$, se obtiene lo siguiente:

$$\begin{aligned}\mathbb{D}_{ij}^2 &= \sum_{k=1}^{d} x_{ik}^2 - 2x_{ik} + x_{jk}^2 \\ &= \sum_{k=1}^{d} (x_{ik} - x_{jk})^2\end{aligned}$$

si tomamos $i = j$, se obtiene la diagonal de $\mathbb{D}^2$ es cero. Por lo tanto, la matriz $\mathbb{D}^2$ es la matriz de distancias entre los vectores ij.

Problema 2.2

En la página 18 del archivo recpat6.pdf de la clase del 9 de febrero, verifica cómo que se obtiene la expresión $K_\Phi(x, y) = (1 + \langle x, y\rangle)^2$. De manera similar, supongamos que se define otro kernel K:

$$K(x, y) = \langle x, y\rangle^3 x, y \in \mathbb{R}^2$$

Busca una función $\Phi()$ tal que:

$$K(x, y) = \langle \Phi(x), \Phi(y)\rangle$$

Sea $X = [x_1, x_2]^T$ y $Y = [y_1, y_2]^T$, calculando $\langle X, Y\rangle^3$ se obtiene lo siguiente:

$$\begin{aligned}
\langle X, Y\rangle^3 &= (x_1y_1 + x_2y_2)^3 \\
&= (x_1y_1)^3 + 3(x_1y_1)^2(x_2y_2) + 3(x_1y_1)(x_2y_2)^2 + (x_2y_2)^3 \\
&= x_1^3y_1^3 + \sqrt{3}x_1^2x_2\sqrt{3}y_1y_2^2 + \sqrt{3}x_1x_2^2\sqrt{3}y_1y_2^2 + x_2^3y_2^3 \\
&= \langle (x_1^3, \sqrt{3}x_1^2x_2, \sqrt{3}x_1x_2^2, x^3), (y_1^3, \sqrt{3}y_1^2y_2, \sqrt{3}y_1y_2^2, y^3)\rangle \\
&= \langle \Phi(x), \Phi(y)\rangle
\end{aligned}$$

por lo tanto:

$$\Phi(z = (z_1, z_2)) = (z_1^3, \sqrt{3}z_1^2z_2, \sqrt{3}z_1z_2^2, z^3)$$

Problema 2.3

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

$$K(A, B) := \#(A \cap B)$$

Busca una función tal que:

$$K(A, B) = \langle \Phi(A), \Phi(B)\rangle$$

Como S es un conjunto finito, entonces podemos decir que el número total de elementos en S es n. Dando asi que $S = \{s_i\}_{i=1}^n$. Sea Φ la siguiente función:

$$\Phi(X) = \sum_i^n \mathbb{I}_X(s_i)$$

donde X es un conjunto (vector) de elementos y $\mathbb{I}_S$ una función indicadora tal que

$$\mathbb{I}_X(s_i) = \begin{cases} 1 & \text{si } s_i \in S \\ 0 & \text{si } s_i \notin S \end{cases}$$

Entonces,

$$\begin{aligned}\langle \Phi(A), \Phi(B)\rangle &= \sum_{i=1}^{n} I_A(s_i) I_B(s_i)\\ &= \sum_{i=1}^{n} I_{A\cap B}\\ &= \#A \cap B\end{aligned}$$

Lo anterior se pudo reducir ya que, la suma dará valores diferentes a cero solo si el elemento s_i se encuentra en los dos conjuntos. Por lo tanto:

$$\Phi(X) = \sum_{i}^{n} \mathbb{I}_X(s_i) \qquad \mathbb{I}_X(s_i) = \begin{cases} 1 & \text{si } s_i \in X \\ 0 & \text{si } s_i \notin X \end{cases}$$

Problema 2.4

Decidimos que dos observaciones x_i , x_j son conectados por una arista en el grafo correspondiente si x_i está entre los k-vecinos más cercanos de x_i o x_j está entre los k-vecinos más cercanos de x_i . Muestra que la adición de una sola observación en este ejemplo puede destruir por completo el desenrollamiento. Márcala en el dibujo y explícalo.

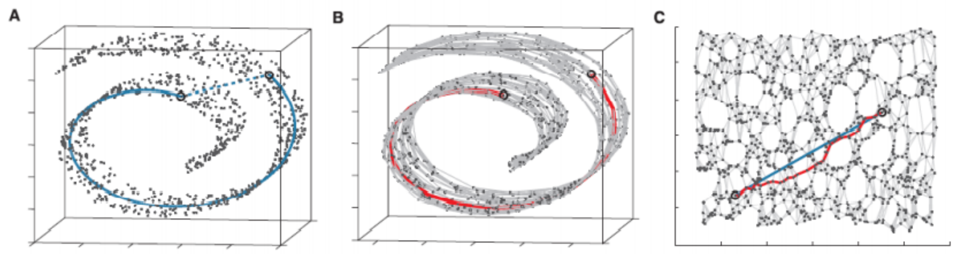


Figura 1: Datos originales dados.

La muestra que se añadiria a los datos mostrados en la figura 1 es un dato entre la sabada formada. Esto puede ilustrarse en la figura 2. Como la figura 1 esta formada con los k-vecinos más cercanos, al añadirel nuevo dato se contemplarían los vecinos del mismo y en que conjuntos estaría involucrado el mismo. Dando así que la figura formada se desenrolle.

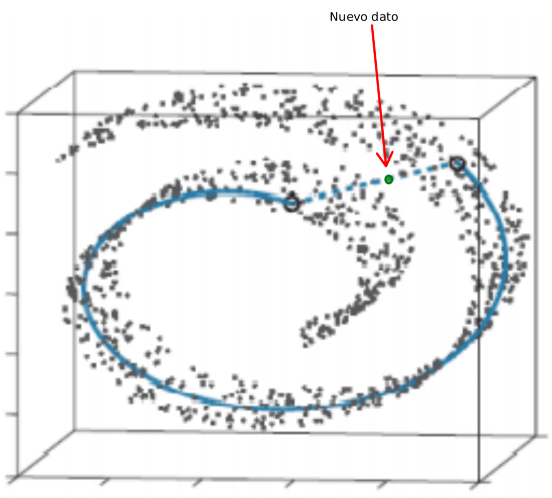


Figura 2: Dato propuesto para desenrollar la figura.

Problema 3.1

Trabajamos con los de datos fashion MNIST. Se trata de imágenes 28x28 de diez diferentes tipos de prendas. Trabajaremos con fashion-mnist_train.csv. Ver https://www.kaggle.com/zalando-research/fashionmnist

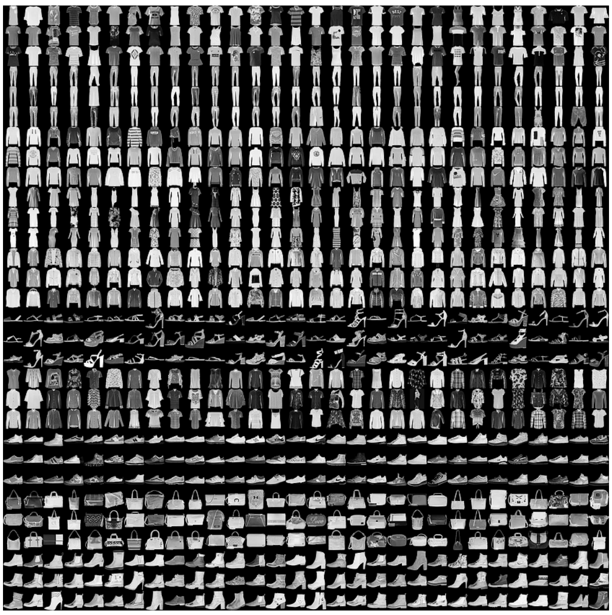
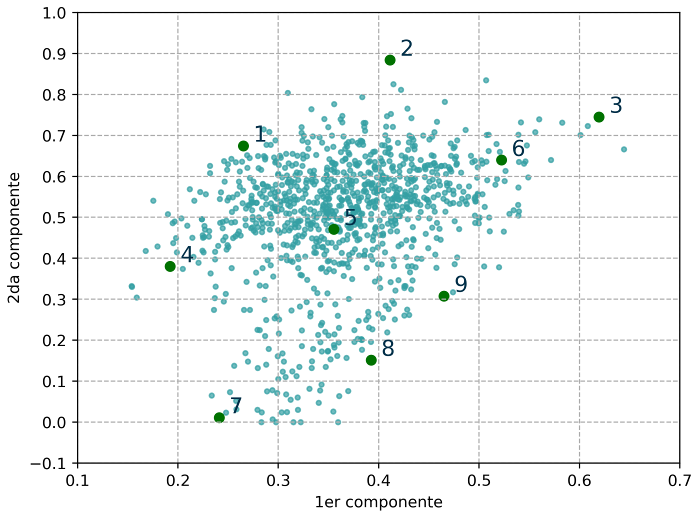


Figura 3

Busca visualizaciones 2D y 3D basadas en PCA de las imágenes de T-shirts (clase "0''). ¿ Ves posible encontrar interpretaciones de los componentes como lo hicimos en clase con la base mnist (clásico) de dígitos?

Visualizaciones 2D



(a)

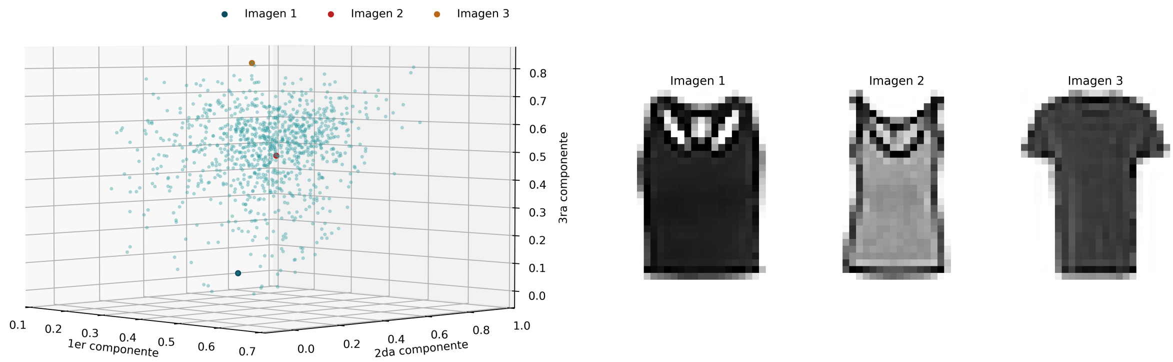
(b)

Figura 4

Imagen	Componente 1	Componente 2
1	0.2652	0.6747
2	0.4111	0.8843
3	0.6193	0.7453
4	0.1923	0.3806
5	0.3552	0.471
6	0.5223	0.64
7	0.241	0.011
8	0.3925	0.1517
9	0.465	0.3078

Tabla 1

Visualizaciones 3D



(a)

(b)

Figura 5

Imagen	Componente		
	1	2	3
1	0.3374	0.4134	0.058
2	0.4574	0.377	0.494
3	0.292	0.5571	0.8176

Tabla 2

Problema 3.2

Trabajamos con datos de calificaciones de peliculas de Netflix por usuarios: `https://grouplens.org/datasets/movielens/latest/` Nos limitamos a la base chiquita. Busca algunas visualizaciones informativas de estos datos y coméntalos. Aplica MDS para obtener una visualización de las películas, explora diferentes kernels (basándose en el vector de calificaciones de cada pelicula y/o los géneros a los cuales cada pelicula pertenece). Hay muchísimas calificaciones faltantes. Limítate a un subconjunto chiquito que se puede trabajar facilmente.

Problema 3.3

¿ Cómo definir una medida de semejanzas entre usuarios/peliculas usando la matriz de calificaciones de peliculas por usuarios?