Trabajo Fin de Máster

20 de febrero de 2018

Resumen

Visualizador...

Índice general

1. Introducción **2**

1.1. Algoritmos lineales 2

1.2. Algoritmos no lineales 5

2. Objetivo **8**

3. Desarrollo **9**

4. Conclusiones **10**

Capítulo 1

Introducción

El universo que habitamos es tetradimensional, es decir, tiene tres dimensiones espaciales y una temporal. Esto nos limita a la hora de poder visualizar objetos (o datos) de mayor dimensión, debido a nuestra incapacidad física de poder imaginar más de tres dimensiones espaciales. Es uso habitual el realizar proyecciones de objetos tridimensionales a bidimensionales.

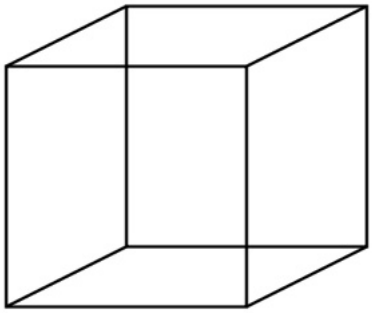

Figura 1.1: Cubo en 2 dimensiones

En Machine Learning la reducción de dimensionalidad es el proceso de reducir el número de variables de un conjunto a un subconjunto de variables manteniendo la mayor cantidad de información posible. La transformación de datos puede ser lineal o no lineal.

1.1. Algoritmos lineales

Estos algoritmos buscan un sistema de referencia de dimensión inferior al original sobre el que proyectar el conjunto de datos. El criterio para encontrar ese

sistema de referencia depende del algoritmo. Entre los más conocidos están:

Principal Component Analysis

Busca la proyección en la que los datos queden representados en términos de mínimos cuadrados. Para ello, se realiza una descomposición en autovalores y se calcula la matriz de covarianza. Los nuevos ejes sobre los que se proyectan los datos son las componentes principales. La primera componente principal esla transformación lineal que maximiza la varianza de los datos, la segunda es la segunda mayor combinación, etc...

Figura 1.2: Principal Component Analysis

En la figura 1.2 se representa el conjunto Iris en tras proyectarlo con el algoritmo PCA en un espacio bidimensional. El conjunto de datos Iris es uno de los datasets más conocidos en la bibliografía. Consta de 150 observaciones de plantas tipo Iris divididas en tres clases (Setosa, Versicolor y Virginica). Para cada una de las

observaciones se dispone de cuatro medidas (variables) de la longitud y anchura de sus sépalos y los pétalos. Se puede observar en la proyección de PCA, las componentes principales (eje x e y) se construyen como componentes lineales de cada una de estas medidas.

Linear Discriminant Analysis

Algoritmo de clasificación que generaliza el discriminante lineal de Fisher. Busca una recta que permita clasificar ambos conjuntos en el espacio. Muy relacionado con PCA (busca las combinaciones de variables que mejor expliquen los datos). Tiene en cuenta las clases a diferencia de PCA.

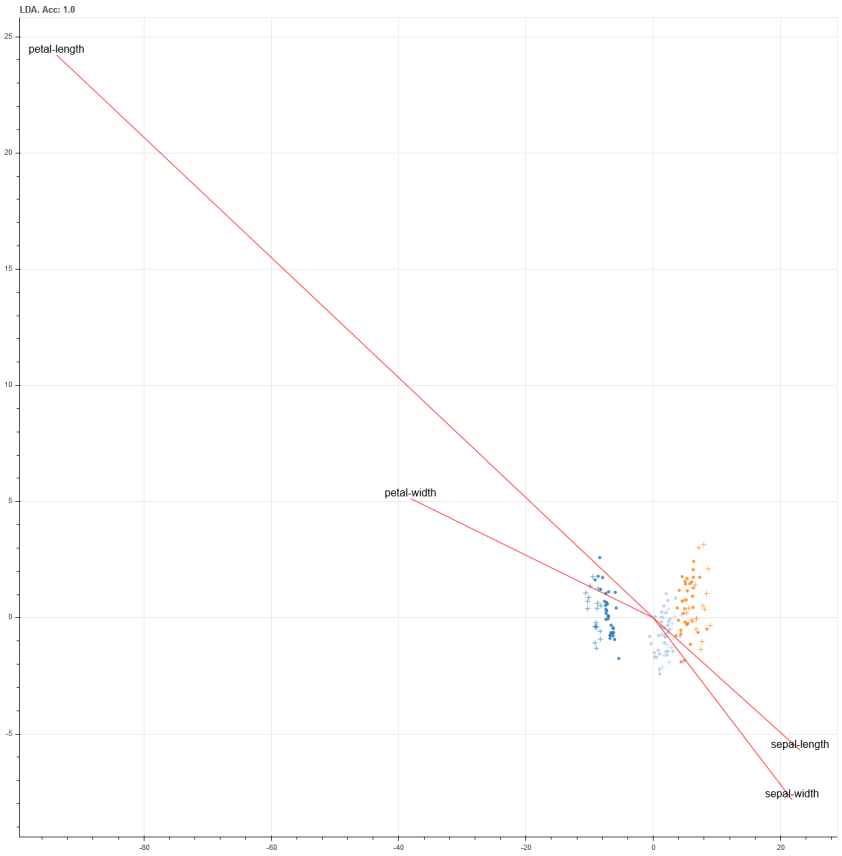

Figura 1.3: Linear Discriminant Analysis

Proyección aleatoria

PCA con un sistema de referencia aleatorio.

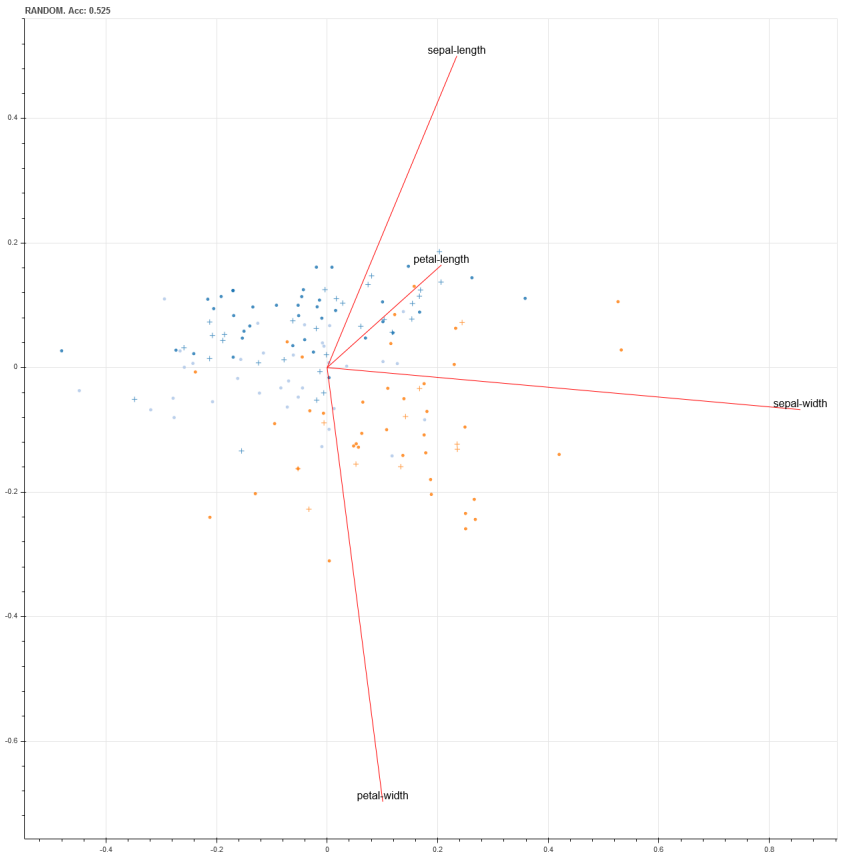

Figura 1.4: Proyección aleatoria

1.2. Algoritmos no lineales

Multi-Dimensional Scaling

Buscar qué factores (dimensiones) subyacen bajo los datos obtenidos en un estudio. Muy utilizado en marketing y ciencias sociales.

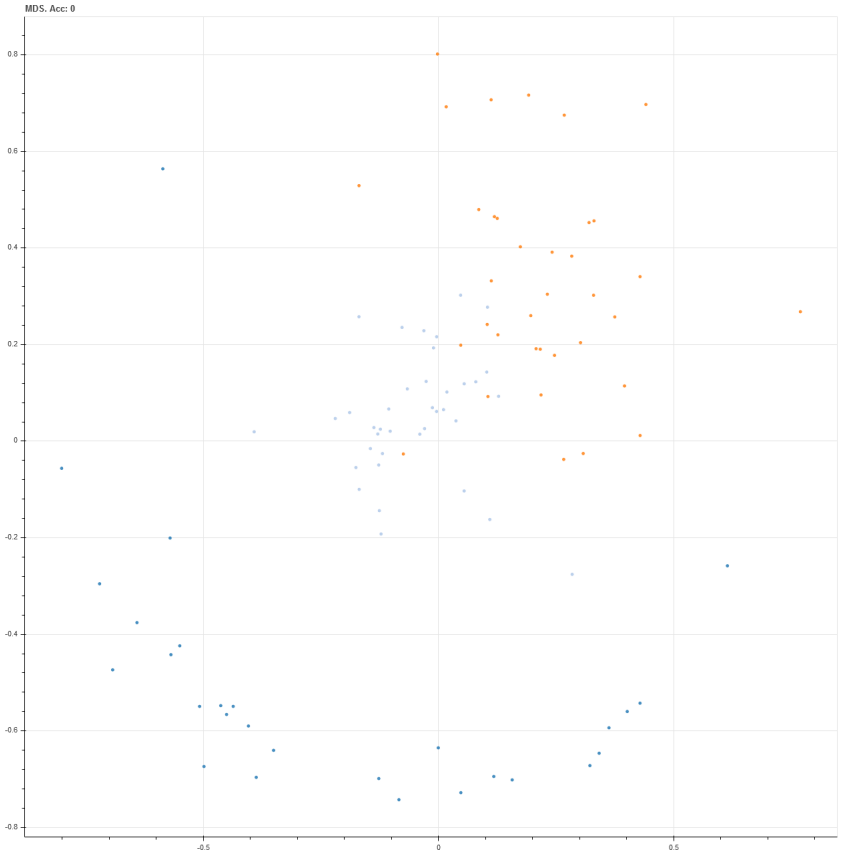

Figura 1.5: Multi-Dimensional Scaling

T-Distributed Stochastic Neighbour Embeding (T-SNE)

Objetos similares en alta dimensión están próximos en el nuevo sistema de referencia.

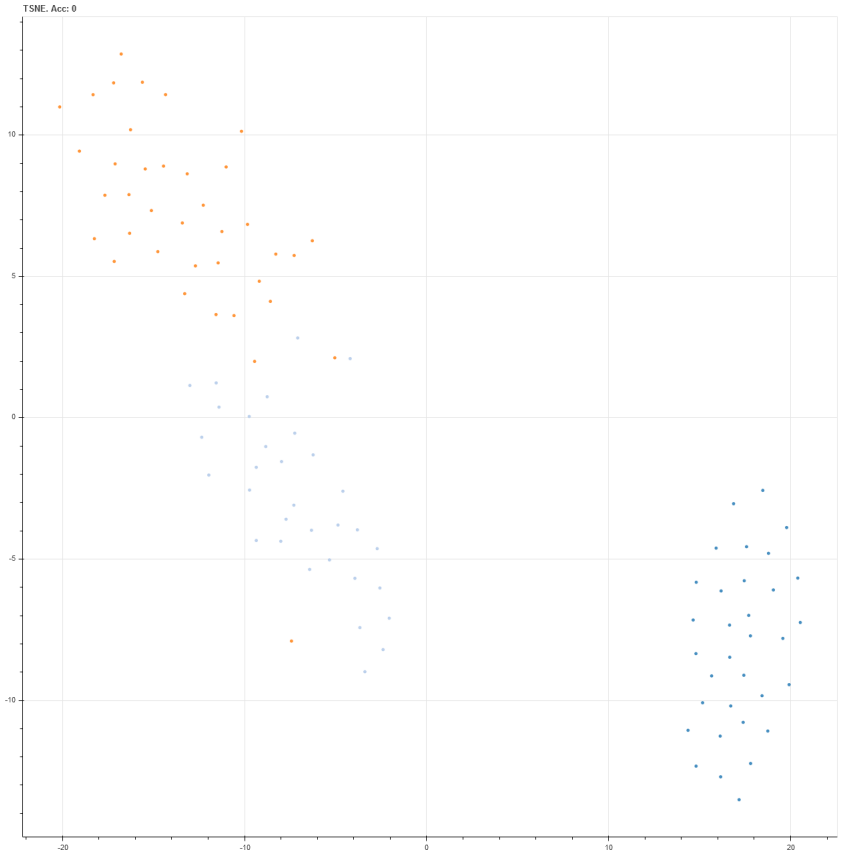

Figura 1.6: T-SNE

Capítulo 2

Objetivo

Capítulo 3

Desarrollo

Capítulo 4

Conclusiones