22-06-2021

Aprendizaje semisupervisado

Establecemos el n de centroides.

Trabaja con las m imágenes que mejor representen la totalidad del conjunto de datos. Selecciona las m imágenes en función de la menor distancia a los centroides establecidos.

Se entrena el modelo con las imágenes más representativas. Con el semisupervisado lo que hago es propagar las clases dependiendo de las distancias a cada centroide. Se asigna a las clases que están entorno a cada una de las imágenes más representativa las que están más cercanas que pueden representar la clase de base o no (si la imagen más representativa es un 4 asignará también un 4 a los más cercanos a ese 4. Tiene un margen de error al asignar a imágenes que no son la representativa (4), pero están cercanas a esa imagen representativa, el valor de esa clase. (por ej. si hay un 5 entre las más cercanas, le asignará también un 4).

Percentile\_closet.- Hay otro algoritmo para que cuando dos clúster están muy cercanos incluya en uno de ellos los valores que están más cerca del centroide más el 20% de distancia. Incluirán estos valores en el clúster con el que se aplique en primer lugar este criterio.

Hay que comprobar posteriormente de forma manual que están correctos los dígitos que se incluyen en cada clúster. Si no es así hay que recolocar a mano el clúster en el que se ha de incluir

TSNE.- algoritmo que reduce dimensiones como PCA. Utiliza el mismo método de proyecciones que PCA

DBSCAN

Genera clúster teniendo en cuenta la densidad de los puntos

eps.- distancia que va a considerar como máximo para que sea un vecino

min\_sample.- establece el mínimo de vecinos que han de existir para considerar que es una instancia central

Tiene en cuenta la magnitud de los datos por lo que si tenemos variables con diferentes magnitudes (rangos):

* previamente tenemos que normalizar/standarizar las variables (standarScaler)
* Aplicar Kmeans que no tiene en cuenta las magnitudes

Dentro del clúster puede haber varias instancias centrales

Ventajas respecto a Kmeans:

* Se pueden detectar los outliers (los -1). Valor -1 no pertenece a ningún clúster. Son las anomalías
* Puede recoger distribuciones

Se pueden mostrar las anomalías para ver qué características tienen

No supervisado:

* Clusterización: Kmeans, DBSCAN
* Para reducir la dimensionalidad: PCA, TSNE