Práctica 4 Introducción al Aprendizaje Automático

Jose Joaquín Rodríguez García y Araceli Teruel Domenech

Universidad Politécnica de Valencia
Master en Big Data Analytics

Octubre 2017

1. Primera Tarea

1.1. Enunciado

A partir del código desarrollado en la práctica anterior, quitar el bucle de validación cruzada y añadir el código para dividir el training set en tantos subconjuntos como clases se sepa existen en el training set.

1.2. Resolución

Quitamos el bucle de validación cruzada y dividimos el training set en tantos subconjuntos como clases se sepa que existen en el training set, en este caso hay 10 subconjuntos

2. Segunda Tarea

2.1. Enunciado

Crear una lista de objetos de la clase sklearn.mixture.GMM para obtener un GMM por cada clase. En principio utilizad 10 componentes por cada GMM. Añadid el código correspondiente para calcular la densidad de probabilidad de una clase a partir de las densidades de probabilidad de cada muestra con respecto a cada componente del GMM de cada clase. p(xn; Ck) = p(Ck) XJ j=1 p(j) p(xn j j) donde J es el número de componentes en el GMM de cada clase, y j es el índice que identifica cada componente. ¿Qué función

de la clase sklearn.mixture.GMM debe utilizarse para obtener las densidades? Mirad el manual de la clase GMM en la documentación del Scikit-Learn. ¿Puede que sea el método score()? Ojo que muchas funciones devuelven el logaritmo de la densidad de probabilidad. ¿Hace falta añadir las probabilidades a priori de cada componente dentro de cada GMM? ¿Dónde está dicha información en el objeto de la clase GMM? ¿O viene ya calculada por la función que utilizamos?

2.2. Resolución

La función que se usa para obtener las densidades es *score()* encontrada en la clase sklearn.mixture.GMM. El método *score()* lo que calcula es el logaritmo de las densidades de cada punto en X. El método *score_samples()* lo que calcula es el logaritmo las probabilidades, por lo que usando esta función no necesitaremos calcularlas previamente.

3. Tercera Tarea

3.1. Enunciado

Una vez resuelta la tarea 2, realizad un barrido para ver el efecto de variar los siguientes hiper-parámetros: el número de componentes de cada GMM el número de componentes para representar cada muestra aplicando PCA, y el tipo de matriz de varianzas-covarianzas. Se debe ir complentando la tabla que se comentó en prácticas anteriores para ver la precisión que puede alcanzarse modificando los hiper-parámetros que afectan al comportamiento de cada técnica.

3.2. Resolución