Algoritmo de mezclas gaussianas o Gaussian Mixture Model (GMM)

En la mayoría de los casos, los datos siguen lo que se denomina una distribución normal o gaussiana, que se puede representar mediante la forma de una campana simétrica respecto al punto medio. En el caso de la clasificación, el conjunto de observaciones puede descomponerse en varias campanas correspondientes a los distintos clústeres que contienen cada una de las observaciones.

Gráfico, Gráfico de superficie

Descripción generada automáticamente

*Curvas gaussianas*

En la figura anterior hemos representado las distintas curvas gaussianas en 3D de nuestro conjunto de observación. Podemos observar una primera curva que representa nuestras cerezas y una segunda un poco menos marcada para nuestros albaricoques.

El programa de esta representación está disponible para su descarga en el sitio del editor.

El algoritmo GMM (mezcla gaussiana) permite determinar los distintos clústeres separando los datos contenidos en las distintas curvas con forma de campana. Es demasiado complejo para este libro y no vamos a dedicar tiempo a explicar su funcionamiento de manera detallada. Sin embargo, sepa que el algoritmo determina una media y una varianza para cada clúster y después determina una probabilidad de pertenecer o no a ese clúster para cada observación.

Para utilizar este algoritmo y visualizar la creación de clústeres efectuada (figura siguiente), le invitamos a introducir estas líneas de programa:

from sklearn import mixture

#Determinación de los clústeres (2 a encontrar)

gmm = mixture.GaussianMixture(n\_components=2)

#Aprendizaje

gmm.fit(frutas)

#Clasificación

clusteres = gmm.predict(frutas)

#Visualización de los clústeres

plt.scatter(frutas.DIAMETRO, frutas.PESO, c=clusteres, s=40,

cmap='viridis');

plt.xlabel("DIÁMETRO")

plt.ylabel("PESO")

plt.show()

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

*Corrección de los errores de las clasificaciones*

De esta manera podemos comprobar que la distribución en clústeres es más conveniente con este algoritmo porque se han corregido los errores de las clasificaciones.

Ahora podemos guardar el modelo y utilizarlo para la predicción, como hemos hecho con el algoritmo de las K-Mean.