Práctica 7. Clustering con KMeans|GMM y GMM como modelo generativo de imágenes

Objetivo

El objetivo de esta práctica es utilizar **KMeans** y **Mezcla de Gaussianas** (**GMM**), para:

1. comprender su uso en problemas de agrupamiento (clustering) y
2. emplear y comprender cómo *GMM* también puede emplearse como estimador de densidad (probabilidad) de los datos, y por tanto, como generador de datos nuevos, mediante el muestreo de esta distribución.

Estudio previo

- Repasa las transparencias de clase y estudia las funciones auxiliares proporcionadas para esta práctica.
- Revisa la documentación de scikit-learn de las clases `KMeans` y `GaussianMixture`.
- Revisa **BIC** como criterio para selección de modelos. En esta práctica será útil para la selección de número de componentes en *GMM*.

Parte 1: Agrupamiento

Empleando datos simulados, vamos a explorar el problema de agrupamiento (*clustering*). Utiliza y sigue el notebook `P7_sinteticos.ipynb` y atiende a estas cuestiones:

a) Comprende cómo se generan los datos simulados. ¿Qué tipo de distribución los está generando? Teniendo en cuenta esta distribución, ¿es *GMM*, en principio, un modelo apropiado para modelar/ajustar a estos datos simulados?

b) Haz y compara el ajuste de *KMeans* y *GMM* a estos datos, empleando el número *correcto* de clústers/componentes Gaussianas. ¿Ves alguna diferencia, por ejemplo, en las asignaciones de los datos a los clústers/componentes Gaussianas? También puedes ver cómo evolucionan los parámetros (centros, medias, covarianzas, etc.) durante la optimización de cada método, a partir de los widgets aportados.

c) Prueba a modificar la forma de inicializar los clústers y componentes Gaussianas, así como el tipo de covarianzas. ¿Hay diferencias entre *KMeans* y *GMM*? ¿Qué método, es, en general, más robusto a peores inicializaciones? ¿Por qué?

d) Selecciona el número de clústers (KMeans) y componentes (GMM) en base a la *distorsión* (en KMeans—ver atributo `inertia_`), y al criterio *BIC* (en GMM—ver método `.bic`). Razona tu selección en ambos casos.

Parte 2: GMM como modelo generativo de imágenes

En esta segunda parte, vamos a usar *GMM* para aproximar la función de densidad de probabilidad correspondiente a la distribución de las imágenes de **Fashion-MNIST**. Muestreando esta distribución de probabilidad, podremos generar imágenes nuevas que no existían hasta ese momento. En otras palabras, vamos a usar *GMM* como un *modelo generativo de imágenes*. El notebook a seguir es `P7_gmm.ipynb`.

Para acelerar los cálculos, vamos a emplear un subconjunto de los datos, así como reducir la dimensionalidad de los mismos con PCA. Cuestiones a atender:

a) Completa la función `gmm_vs_gaussian_components` para: 1) reducir la dimensionalidad del subconjunto de datos sobre los que ajustar los GMMs y 2) visualizar los valores de **BIC** para un número variable de componentes Gaussianas.

b) Usando la función anterior, y para obtener resultados lo más rápido posible: 1) transforma los datos al espacio definido por las 3 componentes más principales, 2) usa un subconjunto de 5000 imágenes, 3) emplea covarianzas de tipo `tied` y 4) razona un número apropiado de componentes Gaussianas a utilizar para el ajuste de la distribución de Fashion-MNIST.

c) Ajusta un GMM con el número de componentes elegido en el apartado anterior y genera nuevas imágenes muestreando puntos de esta distribución. Para visualizar las imágenes muestreadas, recuerda que el GMM se ha ajustado en un espacio de dimensionalidad reducida. Para visualizarlas, os recomendamos usar la función `visualize_subset_in_grid`, definida al inicio del notebook.

d) Con el mismo GMM ajustado, elige una de sus componentes Gaussianas, y samplea/muestrea (y visualiza) imágenes correspondientes a esta componente. ¿El tipo de imágenes generadas es diferente?, ¿por qué?

e) Visualiza en 3D el GMM ajustado empleando la función `plotly_gmm_results_3d` definida al inicio del notebook. ¿En qué espacio se está representando el ajuste?

f) Por último, repite el proceso anterior–desde el apartado b) hasta el e), empleando un número de componentes principales razonable para no perder tanta información. Siguiendo el mismo criterio de antes, ¿ha cambiado el número de componentes Gaussianas a emplear? ¿Por qué?

> **⚠ Warning**
>
> Al representar los datos en un espacio de dimensionalidad mayor, el ajuste/optimización del GMM es significativamente más lento (>1 min. por ajuste). Por ello, os recomendamos analizar el número de componentes Gaussianas (con `gmm_vs_gaussian_components`) en una *celda a parte*, sin código adicional.

g) **Opcional:** Evalúa el impacto que tienen las distintas covarianzas en el ajuste de la distribución (anteriormente solo hemos utilizado `tied`). Cuestiones:

1. ¿Cómo influye el tipo de covarianza en el número de componentes Gaussianas a emplear?
2. Al igual que antes, muestrea la distribución para generar imágenes nuevas, y visualiza el ajuste del GMM en el espacio correspondiente a las 3 componentes más principales. Compara el ajuste de los datos.
3. Si disminuimos el número de imágenes a emplear (`subsample_size`), ¿cómo afecta al BIC (y por ende en la selección del número de componentes Gaussianas)?

A entregar en Moodle

Dos notebook `P7_sinteticos.ipynb` y `P7_gmm.ipynb` con el código de cada apartado, los resultados, su interpretación y las conclusiones que hayas obtenido.

Recuerda:

- Trae la práctica preparada para aprovechar la sesión de prácticas al máximo.
- Si te atascas, pregunta en la sesión o en tutorías.
- Debes citar correctamente todas las fuentes utilizadas.
- Tienes 6 días desde tu sesión para depositar la práctica en Moodle.
- Deberás defenderla en tu próxima sesión de prácticas.