Ciencia de Datos
Tarea 3

Para entregar el 23 de marzo de 2020

1. Este ejercicio es sobre clustering y mezclas de Gaussianas.

   Considera un modelo de mezclas de k distribuciones

   $$f(\mathbf{x}) = \sum_{k=1}^{K} w_k f_k(\mathbf{x}),$$

   donde $w_k \geq 0$ y $\sum_k w_k = 1$. En este caso, supondremos que $f_k = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

   Supón que tienes datos $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n \sim f(\mathbf{x})$, y queremos ajustar el modelo de mezclas de Gaussianas (MMG) para usarlo como un soft-clustering.

   a) Obtén la log-verosimilitud de los datos y los estimadores de máxima verosimilitud para los parámetros del modelo.

   b) Implementa un método de clustering usando el siguiente algoritmo (MMG-EM):

      1) Inicializa los parámetros del modelo y los pesos w_k
      2) *Expectation*: asigna las "responsabilidades" de cada dato, es decir, la asignación de un dato al cluster k, que en este esquema es la probabilidad de que una observación se genere de la distribución k:

      $$\gamma_i^k = P(C(i) = k | X = \mathbf{x}_i) = \frac{w_k f_k(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k w_k f_k(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

      3) *Maximization*: actualiza los parámetros $\boldsymbol{\mu}_k^{\text{new}}$ y $\boldsymbol{\Sigma}_k^{\text{new}}$ usando las responsabilidades obtenidas. Observa que en este paso, usamos la "asignación suave" de cada punto a un cluster k, por lo tanto, cada observación debe ser pesada por su correspondiente responsabilidad, y en consecuencia, el número de puntos "asignados" a algún cluster k será $n_k = \sum_{i=1}^{n} \gamma_i^k$.
      4) Repite los pasos (b) y (c) hasta que la log-verosimilitud converja.

      Prueba tu implementación con un conjunto de datos sintéticos bien escogidos en dos dimensiones y compáralo contra fuzzy $k-$means. Discute los resultados.

   c) Considera el caso en que cada Gaussiana tiene la misma matriz de covarianzas esférica:

   $$\boldsymbol{\Sigma}_k = \sigma^2 \mathbf{I}.$$

   Muestra que, cuando $\sigma^2 \to 0$, el método de MMG-EM y $k-$means coinciden.

2. Implementa kernel $k-$means. Puedes basarte en el paper de Inderjit Dhillon, Yuqiang Guan and Brian Kulis: *A Unified view of Kernel k-means, Spectral Clustering and Graph Cuts. UTCS Technical Report, 2005.* el cual está en la página del curso.

   Escoge (o genera) algunos conjuntos de datos adecuados para verificar la eficiencia y ventajas del método. Comparalo con otros métodos para mostrar en qué casos es mejor su desempeño.

3. Los datos en el archivo `data_fruits_tarea.zip` contienen imágenes preprocesadas de 100×100 pixeles, que corresponden a diferentes tipos de frutas, tomadas en diferentes orientaciones y con diferentes características de forma y maduración. En este ejercicio, tratarás de identificar las frutas obteniendo algunas representaciones a partir de las imágenes (Figura 1).

   a) Obtén una representación de las imágenes en el espacio `RGB` usando la mediana como medida de resumen de los valores en cada canal ¿Puedes identificar patrones interesantes en esta representación?

   b) Realiza PCA y Kernel PCA con un kernel Gaussiano en los datos que obtuviste. ¿Puedes identificar grupos interesantes o informativos de las imágenes en los primeros componentes principales?

   c) Aplica $K-$means y Kernel $K-$means. Verifica si puedes identificar los diferentes grupos de frutas.

   d) Repite los incisos $3b$ y $3c$ usando el espacio `HSV`. Para incluir más información sobre cada dimensión, utiliza la información de los tres cuartiles centrales en cada una de ellas, de forma que tengas una representación en un espacio de tamaño $d = 9$. ¿Notas alguna mejoría?

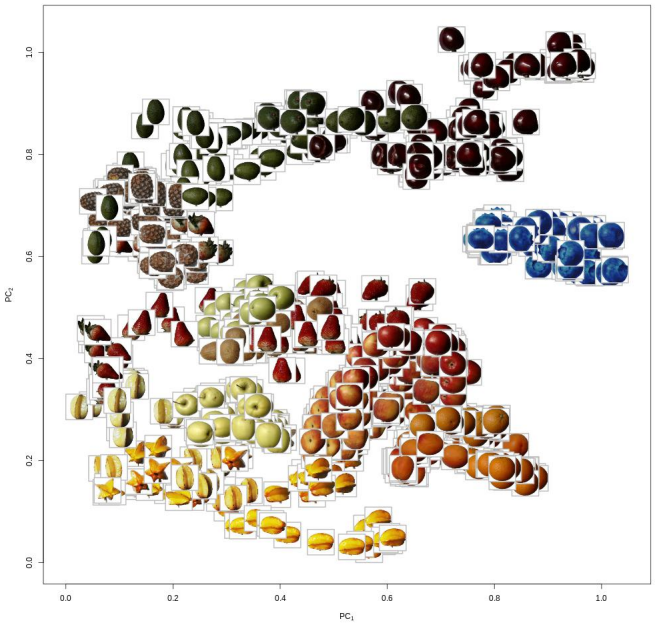

Figura 1: Una representación de las frutas (ejemplo).