Ciencia de Datos
Tarea 3

Para entregar el 14 de marzo de 2019

1. Este ejercicio es sobre clustering y el algoritmo EM.

   Considera un modelo de mezclas de 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.

   - Obtén la log-verosimilitud de los datos y los estimadores de máxima verosimilitud para los parámetros del modelo.
   - Implementa un método de clustering usando el algoritmo EM (MMG-EM) con el siguiente esquema:
     a) Inicializa los parámetros del modelo y los pesos w_k
     b) 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:

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

     c) 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 $n_k = \sum_{i=1}^{n} \gamma_i^k$.
     d) 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.

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

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

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

2. Considera los datos del archivo `gene_expression_2classes.csv`, que contiene la expresión genética de 1000 genes en 40 muestras de tejido, de las cuales las primeras 20 son de pacientes sanos y las 20 restantes de pacientes enfermos de alguna enfermedad.

   a) Compara los métodos de clustering jerárquico y los basados en $k-$means en los datos. ¿Puedes identificar los dos grupos existentes? Prueba usando medidas de disimilaridad euclidena y correlación, así como diferentes tipos de enlace. ¿Cómo cambian los resultados realizando PCA previamente a los datos? ¿Cuántos componentes sugieres usar?

   Realiza un reporte breve de todos estos aspectos, ilustrando de forma adecuada tus hallazgos y conclusiones.

   b) Uno de los médicos quisiera saber qué genes muestran mayores diferencias entre los dos grupos de tejido. ¿Cómo lo verificarias? Implementa tu idea.