

# Sistemas Inteligentes

## Trabajo 5 – Clustering Pt. 2

Profesor: Alejandro Figueroa

Ayudantes: Alexander Espina - Daniel Palomera

Fecha de publicación: 17/08/15

Fecha de entrega: 16/10/2015

### Aspectos generales

- La entrega del informe debe ser realizada, de manera impresa, en la fecha indicada, en la secretaría.
- Se debe utilizar los datos etiquetados por todos los integrantes (máximo dos).

### Objetivos

Fortalecer los conceptos fundamentales de clustering vistos en la tarea anterior, además de introducirlos a Fuzzy C-Means.

### Desarrollo

El alumno debe implementar Fuzzy C-Means en C++. Para ésto, debe definir claramente sus parámetros:

- El número de iteraciones debe ser el mismo que el utilizado en la tarea anterior. De este modo, poder realizar una comparación justa.
- Punto de inicio (inicialización). Misma estrategia que la tarea anterior.

Al igual que K-Means, Fuzzy C-Means utiliza como parámetros el número de clústers a generar. Para el caso del grupo de D. Palomera,  $k$  vale 2, para el grupo de A. Espina,  $k=3$ . Otro parámetro fundamental para Fuzzy C-Means es el epsilon, para este efecto utilicen un valor de 0.00005.

Nótese que las etiquetas **NO** deben ser usada en el espacio vectorial, sino sólo para la verificación de los resultados. El espacio vectorial a utilizar es la bolsa de palabras generadas en tareas anteriores. El parámetro  $m$  debe ser variado

entre 1, ..., 10. Para cada configuración (valor de  $m$ ), calcule la pureza, la entropía, precisión, recall y F1-Score de cada uno de sus clústeres. Posteriormente, calcule el Rand Index, la varianza interna y el índice de Dunn. Para el modelo con el mejor valor de  $m$ , ordene los ejemplos de mayor a menor membresía para cada una de las clases, y grafique la curva respectiva.

Nótese que el informe no sólo contempla el despliegue de los resultados, sino que el alumno debe utilizar una estrategia de presentación de resultados clara y analizar los resultados. Debe interpretar los resultados obtenidos. Para la mejor configuración, puede inspeccionar los vectores centroides, los puntos más cercano a este. Recuerde, que al igual que la gran mayoría de los informes científicos, es altamente deseable analizar errores, es decir puntos asignados a un clúster erróneo. Utilice los resultados de la tarea anterior para contrastar, y validar sus conclusiones.

El alumnos (o su grupo) debe entregar su código. Códigos similares serán vistos como plagio, y recibirán la nota mínima. La tarea se acoge a las normas descritas en el programa del curso.