Sistemas Inteligentes

Trabajo 4 – Clustering

Profesor: Alejandro Figueroa
Ayudantes: Alexander Espina - Daniel Palomera
Fecha de publicación: 17/08/15
Fecha de entrega: 2/10/2015

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

Introducir a los alumnos los conceptos fundamentales de clustering. En especial, K-means, métricas de distancia, y evaluación.

Desarrollo

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

- Número de iteraciones.
- Punto de inicio (inicialización).

K-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 K-Means es la métrica de distancia. El alumno debe mantener los parámetros anteriores fijos, y probar las siguientes métricas de distancia:

- Minkowski (h = 3, 4 y 5)
- Euclideana
- Manhattan

- Euclideana cuadrada
- Chebychev
- Canberra
- Cord al cuadrado
- Chi-squared al cuadrado.

Para las diferentes configuraciones, asuma que la etiqueta de cada clúster es la mayoritaría que contiene. Nótese que las etiquetas **NO** deben ser usada en el espacio vectorial, sino sólo para la verificación de los resultados. El espacio vectorial es la bolsa de palabra utilizada en las tareas anteriores.

Para cada configuración/métrica, calcule la pureza, la entropía, precisión, recall y F1-Score de cada uno de sus clústeres. Posteriormente, para cada configuración, calcule el Rand Index, la varianza interna y el índice de Dunn.

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 analazar errores, es decir puntos asignados a un clúster erróneo.

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.