Universidad Andrés Bello
Facultad de Ingeniería
Ingeniería Civil Informática

Sistemas Inteligentes

Trabajo 3 Aprendizaje No-Supervisado

Profesor: Alejandro Figueroa
alejandro.figueroa@unab.cl
Ayudante: Ayudante: Jean Contreras
j.contrerasleyton@uandresbello.edu
Horario: miércoles 12:10-11:40
Jueves 14:10-15:40

Fecha de Publicación: jueves 7 de septiembre de 2017
Fecha de Entrega: lunes 9 de octubre de 2017
Lugar: Horario de clases/ayudantía/con Haydeé Vidal

Aspectos Generales

- El trabajo es individual.
- La entrega del informe impreso debe ser realizada de manera presencial, en horario de clases.
- Lea atentamente las indicaciones esbozadas en el syllabus del curso.

Objetivos

Fortalecer los conceptos relacionados con aprendizaje no-supervisado: modelos, métricas de evaluación y metodología de trabajo.

Desarrollo

La primera parte de esta tarea consiste en analizar la salida de K-Means. Con este objetivo utilice la representación vectorial de sus datos desarrollada en las tareas anteriores, y ejecute K-means. Para este caso, utilice K como el número de clases que tiene su conjunto de datos. Recuerde que las etiquetas no pueden ser utilizadas para aprender el modelo de K-means. Una implementación de K-means puede ser encontrada en:

http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/yakmo/

Analice los centroides ¿Qué dicen de los datos? Muestre los diez valores más representativos de cada clúster. Utilizando las etiquetas reales calcule la precisión, recall, F1-Score de cada una de las clases. Calcule las entropias y purity respectivos. Analice los resultados.

La segunda parte de esta tarea, consiste en aplica un algoritmo de clustering basado en densidad (DBSCAN). Este algorimo no necesita el número de clústers a crear. Una implementación de DBSCAN puede ser encontrada en:

https://github.com/propanoid/DBSCAN

Haga el mismo análsis para K-means. ¿Qué observa? ¿Hay relación entre los clústeres generado por K-means y DBSCAN? ¿Qué algoritmo es mejor? Comente extensamente argumentando con sus resultados.

Nótese que debe utilizar todo el conjunto de datos como uno solo para esta tarea, es decir no debe hacer cross-validation, ya que los algoritmos no usan las etiquetas.