Universidad de SanAndrés

Aprendizaje No Supervisado – Maestría en Ciencia de Datos
3° trimestre 2022

(Lucas Fernández Piana – fpiana@udesa.edu.ar)

Objetivos de aprendizaje:

El objetivo de este curso es presentar técnicas de aprendizaje no supervisado, es decir que no cuenten con información a priori. Una parte importante del curso será dedicada a los métodos de clustering. El análisis de clusters es una técnica de naturaleza exploratoria que divide los datos en grupos sin utilizar etiquetas previas. En un buen escenario, estos métodos pueden detectar la naturaleza subyacente de la información. Sin embargo, también son útiles como punto de partida para futuros análisis. El primer algoritmo y más popular, K-means, fue propuesto hace más de cincuenta años y desde entonces se han publicado miles de técnicas; lo que demuestra la vigencia del tema. No es caprichoso pensar que esta popularidad se debe a la basta cantidad de aplicaciones que se han desarrollado a partir de esta técnica en disciplinas científicas, en las ciencias sociales, ingeniería, medicina, finanzas, entre muchas otras.

En este curso también abordaremos el problema de reducir la dimensión de los datos. En los últimos años con el avance de la tecnología y la disminución de los costos computacionales hemos vivido una revolución en los sistemas de medición y captura de la información que algunos han llamado "Big Data". Este es el caso de el procesamiento de imágenes, series de tiempo, buscadores de internet, análisis automático de textos, entre muchos otros tópicos. Pero, gran parte de las variables que representan estos datos pueden ser redundantes y llevar a confusiones a la hora del análisis o aumentar exponencialmente el tiempo computacional de un algoritmo. Las técnicas de reducción de dimensión buscan encontrar un subconjunto óptimo de variables o combinaciones de ellas que puedan representar la información sin caer en redundancia.

Durante el curso se dará un enfoque teórico y práctico de los temas. También aprenderemos la matemática detrás de los algoritmos. Se buscará mostrar las aplicaciones con ejemplos de datos reales y simulados. Se darán las herramientas computacionales necesarias, en los dos lenguajes más populares en esta disciplina: R y Python.

Contenidos:

La primera parte del curso introduce el problema de aprendizaje no supervisado comparando con su contraparte supervisada. Se analizan las debilidades de contar con menor información y las expectativas de resultados posibles. Luego, pasaremos a introducir técnicas de clustering que son comúnmente utilizadas y los procesos de validación de resultados que se obtengan. Así mismo, veremos como se puede obtener información sobre las variables utilizando las particiones anteriormente mencionadas.

En una segunda instancia nos enfocaremos en problemas de alta dimensión y en las técnicas de reducción de la dimensión. Veremos que esto tiene implicaciones directas con las técnicas de cluster. Por último, nos enfocaremos en el problema de selección de variables en este contexto.

Carga horaria:

La materia tiene una duración de seis semanas y consta de dos encuentros semanales de tres horas, en cada uno de ellos habrá una clase teórica y otra práctica, cada una de ellas de 90 minutos.

Modalidad de trabajo:

Durante el cursado se integrarán las clases de contenido teórico con la inclusión de actividades prácticas. Se dictarán clases teóricas presenciales (50% de la materia) intercambiando ideas e interpretación con los alumnos e introduciendo actividades prácticas (50% de la materia) para integrar ambos contenidos.

En las clases teóricas se presentarán y desarrollarán los temas de la materia, mientras que en las clases prácticas se hará hincapié en la implementación numérica y recomendaciones específicas sobre las diferentes técnicas y algoritmos. Se incluirán recursos como presentaciones en ppt, artículos de revistas científicas, utilización de diferentes librerías y algoritmos de los softwares utilizados.

La Ejercitación por parte de los alumnos está basada en la resolución de guías de trabajos prácticos con el fin de incorporar paulatinamente los conocimientos esperados. Al analizar problemas con datos reales se hará énfasis en la interpretación de los resultados. Además, la materia tendrá semanalmente un horario destinado a las consultas.

Descripción de las actividades teóricas y prácticas:

Las clases teóricas tendrán una modalidad preponderantemente expositiva, donde se fomentará la discusión y el intercambio de ideas. En las clases prácticas, se darán las herramientas para el correcto uso e implementación de los contenidos dados en las clases teóricas y la adecuada interpretación, alcances y limitaciones de los resultados obtenidos. Orientando a los alumnos a poder realizar las guías de trabajos prácticos.

Mecanismo de evaluación:

El alumno será evaluado de la siguiente manera:
- Examen final multiple choice teórico y examen final práctico.
- Realización de un trabajo final al concluir la cursada que será calificado.
- Asistencia a clases: Será requerida asistencia de por lo menos el 75% de las clases magistrales y prácticas.

Para aprobar el curso es necesario cumplir con las siguientes dos condiciones:
(1) Obtener en el trabajo final una calificación mayor o igual a 4(cuatro) puntos.
(2) Obtener una calificación final mayor o igual a 4 (cuatro) puntos.

Esta calificación final (Cf) será calculado como el promedio ponderado entre la nota del trabajo final (TF) con ponderación 0.3, el exámen final teórico (FT) con ponderación 0.2 y el final práctico (FP) con ponderación 0.5; es decir:

$$Cf = 0.3TF+0.2FT+0.5FP$$

En la calificación final queda a criterio del profesor el redondeo. Si la nota final fuera inferior a 4 puntos, el alumno tendrá la opción de rendir un recuperatorio, que deberá estar aprobado para considerar que la materia está aprobada.

Plagio y deshonestidad intelectual

La Universidad de San Andrés exige un estricto apego a los cánones de honestidad intelectual. La existencia de plagio constituye un grave deshonor, impropio de la vida universitaria. Su configuración no sólo se produce con la existencia de copia literal en los exámenes presenciales, sino toda vez que se advierta un aprovechamiento abusivo del esfuerzo intelectual ajeno. El Código de Ética (http://www.udesa.edu.ar/files/Institucional/Politicas_y_Procedimientos_Universidad_de_San_Andres.pdf) considera conducta punible la apropiación de la labor intelectual ajena, por lo que se recomienda apegarse a los formatos académicos generalmente aceptados (MLA, APA, Chicago, etc.) para las citas y referencias bibliografías (incluyendo los formatos *on-line*). En caso de duda recomendamos consultar el sitio: http://www.udesa.edu.ar/Unidades-Academicas/departamentos-y-escuelas/Humanidades/Prevencion-del-plagio/Que-es-el-plagio. La violación de estas normas dará lugar a sanciones académicas y disciplinarias que van desde el apercibimiento hasta la expulsión de la Universidad.

PROGRAMA

1.- Introducción al problema de clasificación no supervisada. Comparación con el modelo de aprendizaje supervisado. Tipos de métodos de clusters: exclusivos, fuzzy y solapamiento.

2.- Métodos de particiones. K-means y K-medoids. El problema de determinar el número de clusters. Métodos basados en densidades. Algoritmo DBSCAN

3.- Métodos jerárquicos: aglomerativos vs divisivos. Disimilaridad y similaridad. Dendrogramas.

4.- Validación de los clusters. Descartar estructuras aleatorias. Determinar el número de clusters. Validación Externa. Comparaciones entre métodos. Clusters en alta dimensión.

5.- Reducción de dimensión. Casos de estudio. Componentes principales. Aplicaciones a las técnicas de clustering. Multidimensional Scaling. Selección de variables aplicado a clusters y reducción de dimensión.

1. Bibliografía

Aggarwal, C. Reddy, K. (2013), *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC, Data Mining and Knowledge Discovery Series,

Dongkuan Xu, Yingjie Tian, *A Comprehensive Survey of Clustering Algorithms*. Annals of Data Science, Springer, 2014, 2 (2), pp 165–193.

Hastie, T.; Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning* (2nd Edition), Springer, New York Inc., New York, NY, USA.

Jacques, J. y Preda, C. *Functional data clustering: a survey*. Advances in Data Analysis and Classification, Springer Verlag, 2014, 8 (3), pp.24.

James, G**.,** Witten, D., Hastie & T., Tibshirani**,** R. (2013), *An Introduction to Statistical Learning*, Springer New York Inc., New York, NY, USA.

Alboukadel, K. (2017). *Practical Guide To Cluster Analysis in R*. STHDA.

Murphy, K. P. (2012), *Machine Learning: A Probabilistic Perspective*, The MIT Press, Cambrigde, Massachusetts, USA.

Ramsay, J., Silverman, B. W. (1996), *Functional Data Analysis*, Springer Series in Statistics, 1996, New York, NY, USA.

Raftery, A, E y Nema Dean. *Variable Selection for Model-Based Clustering*, Journal of the American Statistical Association, 2012, 101 (473), pp. 168-178.

2. Materiales libres en la web:

a. Software R (http://www.r-project.org/)
b. Softare R (https://www.rstudio.com/)
c. Software Phyton (https://anaconda.org/)
d. Software Python (https://pypi.org/project/pip)