MIAD

Maestría
en Inteligencia
Analítica de Datos

PROGRAMA DEL CURSO

Aprendizaje No Supervisado- 2024-2

Generalidades del curso

Datos generales del curso

- Programa: Maestría en Inteligencia Analítica de Datos
- Nombre del curso: Aprendizaje No Supervisado
- Código del curso: MIID 4303
- Facultad o Departamento: Departamento Ingeniería Industrial
- Periodo académico: 2024-14
- Horario Sesiones Sincrónicas: jueves 6:00-7:50 pm

Equipo docente

Profesor a cargo

- Nombre profesor: Ignacio Sarmiento-Barbieri
  - Correo electrónico: i.sarmiento@uniandes.edu.co

Líder de Tutores

- Nombre: Juan Esteban Segura
  - Correo electrónico: je.segurap@uniandes.edu.co

Tutores

- Nombre: Andrés Felipe Arteta Isaacs
  - Correo electrónico: af.arteta229@uniandes.edu.co
- Nombre: Camilo Bonilla Hernández
  - Correo electrónico: c.bonillah@uniandes.edu.co
- Nombre: Juan Sebastián Paniagua Álvarez
  - Correo electrónico: j.paniagua@uniandes.edu.co
- Nombre: Juan Camilo Prieto
  - Correo electrónico: jc.prietoa@uniandes.edu.co

Horarios de atención a estudiantes

1. Vía el canal de Slack, donde las consultas serán atendidas por los tutores y el profesor en horarios fijos. Estos horarios son:
   a. 12 pm
   b. 9 pm

2. Tutorías virtuales viernes (vía zoom) 7 pm a 9 pm y sábados 10 am a 12pm.

**Recuerda que el canal de comunicación e inquietudes al equipo docente o a la coordinación del programa es a través del correo solicitudes-miad@uniandes.edu.co , los correos personales no son un canal de comunicación oficial*

Introducción al curso

Aprendizaje No Supervisado

Descripción del Curso

El curso es el último en la secuencia de Machine Learning. Se centra en modelos de aprendizaje no supervisado. Estos modelos a diferencia de los vistos en aprendizaje supervisado no cuentan con una respuesta observada. Al no tener una variable de respuesta, no contamos una medida que pueda "supervisar" el análisis. En otras palabras, no supervisado se refiere a que solo se observan predictores, pero no una respuesta asociada.

Así, este curso, complementa los dos cursos anteriores en aplicaciones donde no hay una variable de respuesta definida. Cubre temas de reducción de dimensión, clustering, sistemas de recomendación, y análisis geográfico.

Para evaluar que el estudiante haya cumplido los objetivos planteados, se utilizaran quices y talleres semanales, finalizando con un proyecto grupal de libre aplicación.

En particular el curso busca que el estudiante tenga la capacidad de identificar situaciones reales donde el uso de modelos aprendizaje no supervisado es apropiada, y este preparado para seleccionar, implementar, evaluar, interpretar y comunicar los resultados de estos modelos.

Objetivos de Aprendizaje

Al finalizar el curso el estudiante estará en capacidad de:

- Reconocer las características, usos, ventajas y desventajas de los modelos de aprendizaje no supervisado.
- Identificar situaciones pertinentes para el uso de modelos de aprendizaje no supervisado.
- Seleccionar, crear e implementar modelos de aprendizaje no supervisado apropiados, dependiendo de la disponibilidad y clase de datos.
- Evaluar, comunicar e interpretar los resultados de modelos de aprendizaje no supervisado.

Competencias a desarrollar

- Identificar las particularidades y propósito de diferentes lenguajes de programación, paquetes de software, servicios tecnológicos disponibles en el contexto de análisis de datos y el modelaje matemático.
- Extraer, transformar y cargar datos de diversas fuentes estructuradas y no-estructuradas.
- Resolver modelos de analítica descriptiva, predictiva y prescriptiva.
- Desarrollar o utilizar herramientas que permitan traducir las soluciones de los modelos al contexto de las necesidades de la organización.
- Formular modelos matemáticos a partir de problemas de negocio con el fin de obtener soluciones numéricas que den lugar a análisis que agreguen valor.
- Identificar oportunidades de aplicación de inteligencia analítica para generar valor dentro de las organizaciones.
- Crear narrativas que articulen de manera efectiva los resultados de los modelos analíticos para los stakeholders.

Contenido de la asignatura

El curso está compuesto por 8 semanas que contienen objetivos particulares, un conjunto de lecciones en formato de cuadernos de Jupyter, quices y talleres diseñados para evaluar su conocimiento.

El curso inicia con una motivación sobre el uso de aprendizaje no supervisado y su distinción con los modelos supervisados. Posteriormente en las cuatro primeras semanas se sientan las bases del aprendizaje no supervisado introduciéndose en las dos primeras algoritmos y técnicas básicas de reducción mientras que en la tercer y cuarta semana algoritmos y técnicas para generar agrupaciones o clusters. En las dos siguientes semanas se aplican y expanden estos algoritmos y técnicas para la generación de sistemas de recomendación donde se explora análisis de texto, y análisis de puntos geográficos donde se exploran datos espaciales.

En la tabla a continuación encontrará los objetivos de aprendizaje asociados a cada uno de los módulos o semanas:

Semana	Objetivos
Reducción de Dimensión I: Análisis de Componentes Principales	• Reconocer las características generales del aprendizaje no supervisado y las situaciones en las que es pertinente su uso. • Reconocer el concepto de reducción de dimensión no supervisada y las circunstancias en las que es pertinente aplicarlo. • Reconocer las características y el funcionamiento del análisis de componentes principales. • Construir e implementar el análisis de componentes principales. • Interpretar los resultados del análisis de componentes principales para determinar el número adecuado de componentes a utilizar en la resolución de un problema en particular.
Reducción de Dimensión II: Descomposición en Valores Singulares	• Reconocer las características y el funcionamiento de la descomposición en valores singulares. • Construir e implementar la descomposición en valores singulares. • Interpretar los resultados de la descomposición en valores singulares.
Clustering I: K-medias y K-medoides	• Reconocer en qué situaciones es necesario agrupar datos de forma no supervisada y su importancia. • Reconocer las características y el funcionamiento de los algoritmos de K-medias y K-medoides • Construir e implementar los algoritmos de K-medias y K-medoides • Evaluar e interpretar el desempeño de los algoritmos de K-medias y K-medoides.
Clustering II: Clustering Jerárquico y DBSCAN	• Reconocer las características y el funcionamiento de algoritmos de clustering jerárquico y DBSCAN. • Construir e implementar los algoritmos de clustering jerárquico y DBSCAN. • Evaluar e interpretar el desempeño de los algoritmos de clustering jerárquico y DBSCAN. • Definir el conjunto de algoritmos más apropiado en la resolución de un problema en particular.
Sistemas de	• Reconocer las características generales de los Sistemas de

Recomendación I	Recomendación. • Reconocer las características y el funcionamiento de algoritmos de filtrado colaborativo basado en usuarios. • Reconocer las características y el funcionamiento del algoritmo de Análisis de Canasta de Compra. • Crear e implementar modelos de filtrado colaborativo basado en usuarios y de Análisis de Canasta de compra. • Evaluar e interpretar los resultados de filtrado colaborativo basado en usuarios y de Análisis de Canasta de compra.
Sistemas de Recomendación II	• Reconocer las características y el funcionamiento de las expresiones regulares • Reconocer las características y el funcionamiento de algoritmos de filtrado colaborativo basado en contenido. • Reconocer las características de los modelos de tópicos • Implementar expresiones regulares. • Crear e implementar modelos de filtrado colaborativo basado en contenido y modelo de asignación latente de Dirichlet. • Evaluar e interpretar los resultados de filtrado colaborativo basado en contenido y modelo de asignación latente de Dirichlet.
Análisis de Puntos Calientes Geográficos I	• Reconocer el concepto de puntos calientes, estimación de histogramas y estimación de densidad de kernel. • Identificar y organizar datos georreferenciados. • Crear e implementar modelos de estimación de densidad kernel bivariada para puntos calientes. • Interpretar los resultados de modelos de estimación de densidad kernel bivariada para determinar puntos calientes. • Construir e implementar el conjunto de algoritmos más apropiado en la resolución de un problema en particular.
Análisis de Puntos Calientes Geográficos II	• Reconocer las características y el funcionamiento del modelo de mezclas gaussianas • Crear e implementar modelos de mezclas gaussianas para el análisis de puntos calientes y clustering. • Evaluar e interpretar el desempeño y los resultados de modelos de mezclas gaussianas para el análisis de puntos calientes y clustering con datos georreferenciados.

Metodología

El curso contiene distintas actividades diseñadas para cumplir con los objetivos de aprendizaje generales del curso y particulares propuestos en cada una de las semanas. De esta forma cada semana contiene videos, lecturas interactivas, talleres y quices que le permitirán abordar y aprender.

El curso cuenta también con un proyecto grupal transversal. En este proyecto los estudiantes pondrán en práctica la identificación e implementación de modelos de aprendizaje no supervisado para resolver un problema particular de su elección.

Este es un curso de tres créditos, y por lo tanto exige una dedicación estimada de 18 horas semanales. Sin embargo, este puede variar y el tiempo reflejado en las actividades en plataforma es menor y estimativo. El tiempo de plataforma, no contempla el tiempo que los estudiantes pueden dedicarle a tomar notas, repasar, y/o entender el contenido.

Herramientas y requerimientos tecnológicos

El curso utilizará:

- Jupyter notebooks – Python (tutorial en el curso).
- Git and GitHub (tutorial en el curso).

Conocimientos previos requeridos para tomar el curso

- Conocimientos básicos en probabilidad y estadística.
- Conocimientos sobre modelos lineales estadísticos.
- Conocimientos básicos en cálculo.
- Conocimientos básicos en álgebra lineal.
- Conocimientos básicos de programación en Python.
- Nociones de optimización.
- Conocimientos básicos de Machine Learning comparables a los dictados en Introducción al Machine Learning y Machine Learning y Procesamiento de Lenguaje Natural.

Relevancia del curso y conexión con los demás cursos del programa

- *TRAYECTORIA 1*
    - *Decision analysis:* provee parte conceptual sobre como estructurar un proyecto de análisis.
    - *Laboratorio computacional de analytics:* proporciona los conocimientos básicos de programación, especialmente los asociados a Python que es el lenguaje que se utiliza en el curso.
    - *Modelos de análisis estadístico:* es prerrequisito y da al alumno conocimientos básicos sobre conceptos necesarios en el curso: algebra lineal, medidas de centralidad,

distribuciones de probabilidad.

- *Optimización:* Este curso es prerrequisito y provee al alumno con conceptos fundamentales para el desarrollo del curso, especialmente a los que se refiere a estructuras de problemas de optimización, su naturaleza, y el rol de los parámetros de ajuste.

- *TRAYECTORIA 2*
  - *Introducción al Machine Learning:* es prerrequisito y en el mismo se desarrollan los conceptos básicos de problemas de machine learning.
  - *Machine Learning y procesamiento del lenguaje natural:* es prerrequisito, y sienta bases de problemas que luego serán abordados desde una perspectiva no supervisada.

Criterios de evaluación y aspectos académicos

Para evaluar el desempeño y evolución de aprendizaje el curso cuenta con actividades semanales. El aporte de cada una de las actividades calificables en la nota definitiva del curso son los siguientes:

- Evaluación individual:
  - Quices semanales teórico-prácticos, la calificación más baja será descartada (28%, 4% c/u)
  - Casos-talleres individuales (12%)
    - Semanas 2, 4, 6 y 8 calificados por pares (3 %, c/u)
- Evaluación grupal:
  - Casos-talleres grupales (Sumativo, heteroevaluación) (28%)
    - Semanas 1, 3, 5 y 7 calificados por tutores (7%, c/u)
  - Proyecto libre guiado por los tutores con dos entregas (Sumativo, heteroevaluación) (30%)
    - Entrega 1 semana 4 (14%)
    - Entrega 2 en semana 7 (18%)

Las actividades grupales deberán realizarse en equipos entre **tres y cinco** personas. Los estudiantes son libres de conformarlos y decidir con quienes quieren trabajar. La conformación del equipo deberá comunicarse a través de la plataforma antes de la fecha allí estipulada. **Los estudiantes que para esta fecha no hagan parte de ningún equipo, serán asignados aleatoriamente a equipos de trabajo.**

Para los casos-talleres individuales de las semanas 2, 4, 6 y 8 el estudiante es libre de trabajar en equipo, **pero es responsable de realizar y entregar su propia versión del trabajo asignado**. Esto significa que, a pesar de colaborar y discutir los temas con los compañeros de grupo, la entrega debe ser realizada **individualmente**. **Está permitido compartir código** entre los compañeros de grupo para fines educativos y para facilitar la comprensión del problema a resolver, pero el análisis y las conclusiones deben ser propias. La universidad toma muy en serio el fraude académico y no será tolerado.

Consideraciones

- Fechas importantes: consultar el portal de registro.
- Parámetros de calificación de actividades académicas:
    - Las actividades serán calificadas de forma automática por la plataforma de Coursera. Sin embargo, algunas prácticas computacionales tendrán retroalimentación por parte de los tutores del curso.
    - No se recibirán actividades por fuera de la fecha establecida.
- Actividades grupales: los grupos serán de tres (3) o cuatro (4) estudiantes. Los grupos serán seleccionados por los estudiantes.
- Reclamos: Las solicitudes de revisión de nota se deben hacer por escrito vía Salesforce, respetando el máximo plazo establecido por el reglamento general de estudiantes de maestría (Art. 62).
- Política de aproximación de notas:
    - Para aprobar el curso la nota ponderada total debe ser superior o igual a 3.00.
    - La nota definitiva del curso se aproximará a 2 decimales dentro de la escala numérica entre 1.50 y 5.00. En caso de no cumplir la regla anterior, la nota definitiva será el mínimo entre 1.50 y la nota aproximada a dos decimales.

**Recuerda que, de acuerdo al nuevo reglamento de estudiantes de maestrías, todo estudiante que desee formular un reclamo sobre la calificación de cualquier evaluación o sobre la nota definitiva del curso deberá dirigirlo por escrito y debidamente sustentado al profesor responsable de la materia, dentro de los cuatro (4) días hábiles siguientes a aquel en que se dan a conocer las calificaciones en cuestión. El profesor dispone de cinco (5) días hábiles para resolver el reclamo formulado.*

Bibliografía

- Ahumada, H. A., Gabrielli, M. F., Herrera Gomez, M. H., & Sosa Escudero, W. (2018). Una nueva econometría: Automatización, big data, econometría espacial y estructural.
- Amat Rodrigo, Joaquín (2022) Ajuste de distribuciones con kernel density estimation y Python. Disponible en https://www.cienciadedatos.net/documentos/pystats02-ajuste-distribuciones-kde-python.html. Accedido el 10 de Abril de 2022
- Amat Rodrigo, Joaquín (2022). Clustering con Python. Available under Attribution 4.0 International (CC BY 4.0) at https://www.cienciadedatos.net/documentos/py20-clustering-con-python.html. (Accedido 9 de enero 2022)
- Amat Rodrigo, Joaquín (2022). Clustering y heatmaps: aprendizaje no supervisado. Available under Attribution 4.0 International (CC BY 4.0) at https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps. (Accedido 9 de Enero 2022)
- Amat Rodrigo, J. Detección de anomalías con Gaussian Mixture Models (GMM) y python. Disponible en https://www.cienciadedatos.net/documentos/py23-deteccion-anomalias-gmm-python.html. Accedido el 13 de Abril 04 de 2022
- Amat Rodrigo, Joaquín. (2018). Reglas de asociación y algoritmo Apriori con R, available under a Attribution 4.0 International (CC BY 4.0) at https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion. Accedido el 12 de Enero de 2022
- Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. Applied Statistics, 39, 357–365. doi: 10.2307/2347385.
- Banik, R. (2018). Hands-on recommendation systems with Python: start building powerful and personalized, recommendation engines with Python. Packt Publishing Ltd.
- Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".
- Bishop, Christopher M. Pattern Recognition and Machine Learning. (2006). Springer-Verlag Berlin, Heidelberg.
- Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R (Vol. 747248717, pp. 237-268). New York: Springer.
- Cottam, J., Lumsdaine, A., & Wang, P. (2014). Abstract rendering: Out-of-core rendering for information visualization. Proceedings of SPIE The International Society for Optical Engineering. 9017. 90170K. 10.1117/12.2041200
- Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube

recommendations. Proceedings of the 10th ACM Conference on Recommender Systems. https://doi.org/10.1145/2959100.2959190

- DANE (29 de septiembre de 2020). Encuesta nacional de presupuestos de los hogares (ENPH). Anexos: 32 ciudades y 6 ciudades intermedias. <https://www.dane.gov.co/files/investigaciones/boletines/enph/ciudades-enph-2017.xls>
- Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). Mathematics for machine learning. Cambridge University Press.
- Dempster, Arthur P., Laird, Nan M., and Rubin, Donald B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, 39(1), 1–38.
- Dougherty, J., & Ilyankou, I. (2021). "Hands-On Data Visualization". O'Reilly Media, Inc.
- Ester, M., H. P. Kriegel, J. Sander, and X. Xu. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226–231.
- ESRI, Environmental Systems Research Institute. (1998). "ESRI Shapefile Technical Description." Disponible en: https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf. Accedido el 4 de febrero de 2022
- Fradejas Rueda, J. M. (2020). Cuentapalabras. Estilometrıa y análisis de texto con R para filólogos.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
- Google developers. (n.d.). Recommendation systems. Google. Accedido el 3 de Abril de 2022. Disponible en https://developers.google.com/machine-learning/recommendation/overview
- Google developers. (n.d.). Embeddings: Motivation From Collaborative Filtering. Accedido el 3 de Abril de 2022. Disponible en https://developers.google.com/machine-learning/crash-course/embeddings/motivation-from-collaborative-filtering
- Harrington, Peter (2012). Machine learning in action. Simon and Schuster.
- Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.
- Hoffman, M., Bach, F., & Blei, D. (2010). Online learning for latent dirichlet allocation. advances in neural information processing systems, 23.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions. Series A, Mathematical, physical, and engineering

sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

- Jones, Aaron; Kruger, Christopher; Johnston, Benjamin. The Unsupervised Learning Workshop: Get started with unsupervised learning algorithms and simplify your unorganized data to help make future predictions. Packt Publishing. Kindle Edition.
- Jurafsky, D., &; Martin, J. H. (2020). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Pearson.
- Kaufman, L., &; Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis. Wiley.
- Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of Medoids. Statistical data analysis based on the L1–norm and related methods, edited by Y. Dodge.
- Li, Q., and Racine, J. S. (2007). Nonparametric econometrics: theory and practice. Princeton University Press.
- Macnaughton Smith, P., Williams, W., Dale, M. & Mockett, L. (1965). Dissimilarity analysis: a new technique of hierarchical subdivision, Nature 202: 1034–1035.
- MacWright, T. "lon lat lon lat". Disponible en https://macwright.com/lonlat/. Accedido el 4 de febrero de 2022
- Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.
- Pagan, A. and Ullah (1999). Nonparametric Econometrics. Cambridge University Press.
- Patel, A. A. (2019). Hands-on unsupervised learning using Python: how to build applied machine learning solutions from unlabeled data. O'Reilly Media.
- Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), pp.2825–2830.
- Peña, D. (2002). Análisis de datos multivariantes (Vol. 24). Madrid: McGraw-Hill.
- Pérez López, C. (2004). Técnicas de análisis multivariante de datos. Aplicaciones con SPSS, Madrid, Universidad Complutense de Madrid, 121-154.
- Rahmah, N., & Sitanggang, I. S. (2016). Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra. In IOP Conference Series: Earth and Environmental Science (Vol. 31, p. 012012). IOP Publishing. https://doi.org/10.1088/1755-1315/31/1/012012
- Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modeling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).
- Rey, S., & Arribas-Bel, D. (nd). Introduction Geographic Data Science with PySAL and the pydata stack
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In

Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408).

- Sander, J., Ester, M., Kriegel, HP. et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Mining and Knowledge Discovery 2, 169–194 (1998). https://doi.org/10.1023/A:1009745219419
- Silverman (1998). Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC.
- Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. In ACM Transactions on Database Systems (TODS), 42(3), 19.
- Taddy, Matt; Taddy, Matt. Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions (p. 202). McGraw-Hill Education.
- Tenkanen, H. "Spatial data science for sustainable development". Disponible en https://sustainability-gis.readthedocs.io/en/latest/index.html. Accedido el 4 de febrero de 2022
- Tsai, K. T. (2021). Machine Learning for Knowledge Discovery with R: Methodologies for Modeling, Inference, and Prediction.
- VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. O'Reilly Media, Inc.
- Weizenbaum, J. 1966. ELIZA – A computer program for the study of natural language communication between man and machine. CACM, 9(1):36–45.