

INFERENCIA ESTADÍSTICA Y RECONOCIMIENTO DE PATRONES

Carrera/s: Licenciatura en Ciencia de Datos / Tecnicatura en Programación

01 – Cantidad de horas semanales y totales

Carga Horaria Total: 96 horas

- Teoría: 48 horas
- Práctica: 48 horas

Carga Horaria Semanal: 6 horas

02 – Nombres de las/los integrantes del equipo docente

Docentes:

- Sebastián Simondi
- Rui Marques Rojo

03 – Fundamentación

Al llegar a este curso los estudiantes ya poseen conocimientos básicos de Análisis Matemático, Álgebra y Probabilidades y Estadística. A partir de esta base Inferencia Estadística y Reconocimiento de patrones es una asignatura que se propone introducir a los y las estudiantes a las técnicas básicas, algorítmicas y estadísticas, de Aprendizaje Automático (Machine Learning) para clasificación, supervisada y no supervisada, de datos.

Además, y como parte indispensable para la clasificación, los estudiantes adquirirán también técnicas básicas de reducción de la dimensión, herramienta indispensable a la hora de visualizar e interpretar datos, y de reconocer patrones de forma exploratoria.

Tanto las técnicas de Aprendizaje Automático para clasificación, como las de reducción de la dimensión, son herramientas fundamentales para el desarrollo profesional en la Ciencia de Datos, y sientan las bases para continuar con otras técnicas y modelos de Inteligencia Artificial más complejos.

Otro aspecto importante de esta asignatura es ensamblar la parte teórica de las técnicas antes mencionadas, con su aspecto práctico, pues sendas partes presentan dificultades de índole diferente. Para la Ciencia de Datos es indispensable manejar correctamente tanto la parte teórica como la parte práctica.

La asignatura Inferencia Estadística y Reconocimiento de Patrones proveerá a los y las estudiantes del ciclo completo del análisis de datos, desde su primer acercamiento exploratorio, hasta su análisis más teórico, posterior implementación práctica y extracción de conclusiones.

Finalmente se hará hincapié en la confección de un informe de carácter profesional, volcando allí toda la información pertinente.

04 – Programa sintético

Introducción al problema de clasificación. Introducción a la maldición de la dimensión. Análisis de Componentes Principales.

Introducción al problema de clasificación supervisada. Clasificador bayesiano ingenuo. Análisis del discriminante lineal y cuadrático. Máquinas de soporte vectorial. Regresión Logística. Sobreajuste y métricas de evaluación de modelos y técnicas.

Introducción al problema de clasificación no supervisada. K-medias. Mezcla de gaussianas. K vecinos más cercanos. DBSCAN y OPTICS. Agrupamiento jerárquico. Métricas de evaluación de agrupamientos.

05 – Objetivos

Al finalizar el curso el o la estudiante estará en capacidad de:

- Abordar un conjunto de datos mediante reducción de la dimensión, para un primer reconocimiento exploratorio de patrones.
- Identificar cuándo un problema es de clasificación supervisada, y cuándo es de clasificación no supervisada.
- Evaluar las diferentes técnicas de clasificación supervisada, de acuerdo a la naturaleza del objetivo propuesto.
- Evaluar las diferentes técnicas de clasificación no supervisada, de acuerdo a la naturaleza del objetivo propuesto.
- Implementar las técnicas de reducción de la dimensión, y de clasificación supervisada y no supervisada.
- Extraer valor agregado de los datos utilizando la clasificación como herramienta.
- Extraer conclusiones de los resultados provenientes de aplicar las diferentes técnicas de clasificación, fundamentando adecuadamente.
- Realizar un informe de carácter profesional sobre el problema de clasificación que se haya propuesto resolver.

06 – Propósitos de la enseñanza

El curso es teórico-practico, donde el estudiante debe desarrollar problemas en grupo e individualmente, que permita poner en práctica los conocimientos adquiridos.

Las prácticas serán definidas por parte del profesor el cual dará únicamente los lineamientos generales a seguir. El o los problemas escogidos y su solución aplicada deben ser volcados a un informe y expuestos por los grupos respectivos.

07 – Contenidos

Los contenidos del programa estarán resumidos en los siguientes módulos:

- Introducción al problema de clasificación: aspectos algorítmicos y estadísticos.
- Reducción de la dimensión: introducción al problema mediante la maldición de la dimensión, y posible solución por análisis de componentes principales (ACP). Implementación en Python de ACP.
- Clasificación supervisada: introducción al problema, clasificador bayesiano ingenuo, análisis de discriminante lineal y cuadrático, regresión logística. Sobreajuste y métricas de evaluación de modelos. Implementación en Python de todas las técnicas y métricas.
- Clasificación no supervisada: introducción al problema, k-medias, mezcla de gaussianas, DBSCAN y OPTICS, agrupamiento jerárquico. Métricas de evaluación de agrupamientos. Implementación en Python de todas las técnicas y métricas.

08 – Bibliografía y recursos audiovisuales

Bibliografía Obligatoria:

- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.
- Gareth, J.; Witten, D.; Hastie, T.; Tibshirani, R. (2021). *An Introduction to Statistical Learning*. 2nd ed.

Bibliografía Complementaria:

- Hastie, T.; Tibshirani, R.; Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2nd. ed.
- Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press: London. Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin. (2012). *Learning from Data, a short course*. AML: USA.

09 – Metodología

La cátedra estará compuesta por clases teórico-prácticas virtuales “presenciales”, ejercitaciones y trabajos prácticos grupales.

Las clases teórico-prácticas estarán divididas del siguiente modo:

- Exposición de los conceptos teóricos y debate sobre los temas dados.
- Ejercitaciones para fijar los contenidos.

Trabajos prácticos:

- Desarrollo de trabajos prácticos de investigación.
- Trabajos individuales y grupales
- Seguimiento de trabajos prácticos desarrollados en lenguaje Python.

10 – Uso del campus virtual e integración de TIC en la propuesta pedagógica

La transformación de la educación durante los últimos tiempos nos apremia a poner foco y maximizar el uso de TIC's para implementar la enseñanza "virtual" o "a distancia", preparándonos para la educación "post-pandemia".

El Campus Virtual será el portal de conocimiento en donde los alumnos podrán la propuesta educativa y sus recursos. Complementando su uso, se usarán otras tecnologías para mejorar la participación e involucramiento de los estudiantes, tales como, clases video-presenciales, herramientas online para la resolución de problemas, y recursos mediáticos para reforzar la propuesta educativa.

Finalmente, se utilizará el medio para continuar y reforzar el trabajo de los alumnos en grupos, ya que sin un "aula física" la formación de "grupos" e interacciones entre los estudiantes es reducida.

11 – Evaluación

La evaluación corresponderá a una nota global entre:

- Entrega de ejercicios.
- Dos trabajos prácticos (con derecho a recuperatorio)

12 – Cronograma de actividades / Planificación de clases

Semana	Temas
04-Sep	Introducción al problema de clasificación.
11-Sep	Análisis de Componentes Principales.
18-Sep	Clasificador bayesiano ingenuo.
25-Sep	Análisis del discriminante lineal y cuadrático.
02-Oct	Regresión Logística
09-Oct	Sobreaajuste. Métricas de evaluación de modelos.
16-Oct	Consultas primer trabajo práctico, y semana de entrega del mismo.
23-Oct	Introducción al problema de clustering.
30-Oct	K-medias. Mezcla de gaussianas.
06-Nov	DBSCAN y OPTICS.
13-Nov	Clustering jerárquico. Métricas de evaluación de modelos de clustering.
20-Nov	Consultas segundo trabajo práctico y semana de entrega del mismo.
27-Nov	Consultas recuperatorios y semana para los mismos.