IIC2433 — Mineria de Datos

Programa de Curso

Página de Curso

Importante: la página del curso está en `https://github.com/IIC2433/Syllabus-2021-2-`. Ahí se puede acceder a la información de contacto del profesor y los ayudantes.

Descripción

El desarrollo de la tecnología ha hecho que la mayoría de los datos almacenados de forma física ahora lo estén de forma digital. Esto hace que podamos extraer información de estos datos, mediante algoritmos computacionales, ya sea patrones, modelos de predicción o identificar anomalías. En este curso se ve una batería de técnicas para poder lidiar con información mediante algoritmos computacionales, desde el manejo de datos, construcción de clasificadores o predictores, clusterización, hasta técnicas para la transformación de datos orientada a trabajo con datos semi-estructurados o multidimensionales.

Objetivo General

El objetivo de este curso es proporcionar al alumno elementos que le permitan entender las principales teorías y practicas en Minería de Datos. Al final del curso, el alumno podrá aplicar las principales técnicas utilizadas en la creación de programas capaces de extraer conocimiento desde distintas fuentes y distintos tipos de datos. Además, el alumno contará con fundamentos teóricos para poder decidir qué herramienta aplicar, conociendo sus potencialidades y limitaciones. Finalmente, los alumnos podrán vivir una experiencia real de aplicación de estas herramientas en un entorno realista.

Metodología

El curso se reúne dos veces a la semana. La clase es los días martes, de 14:00 a 16:50, y los jueves de 14:00 a 15:20 habrá un horario de trabajo y consulta. La clase, en si misma, se divide en dos bloques, uno para ver contenidos teóricos y el otro para realizar un laboratorio práctico. Es decir, semana a semana, las alumnas tendrán el siguiente esquema:

- Martes a las 14:00: clase expositiva
- Martes, después de la parte expositiva: laboratorio, trabajo práctico.
- Jueves de 14:00 a 15:30: ayudantía, trabajo de laboratorio y consulta.

Contenidos

1. **Modelos de predicción, clasificación, explicación**
   a) Regresión lineal y logística
   b) Métodos de afinamiento y testeo
   c) Vecinos más cercanos
   d) Árboles y bosques de decisión
   e) Técnicas bayesianas, inferencia causal
2. **Clustering**
   a) K-means, DBSCAN
   b) Modelos de Gaussian Mixtures
   c) evaluación

3. **Preparación y Transformacion de la información**
   a) Pandas
   b) Análisis de componentes principales
   c) Método de kernel
   d) Autoencoders
   e) Encoders word-2-vec, grah-2-vec.
4. **Análisis de información semi-estructurada**
   a) Texto
   b) Grafos

Evaluación

El curso contempla 5 tareas de programación individual, más una tarea optativa, más un proyecto final de grupos de a cuatro personas. Existen, además, una serie de actividades que tienen caracter formativo, pero cuya compleción obtendrá un beneficio en la nota final del curso. **Importante:** La situación sanitaria podría provocar una reducción en el número de evaluaciones.

Tareas. Las tareas se publican los martes, en horario de clases, y deberán ser entregadas el viernes de la semana siguiente a su publicación. Consisten en ejercicios de programación individual, ya sea de análisis de datos o de implementación de algoritmos.

Proyecto. El proyecto consta de tres etapas: Planificación, Avances y Entrega Final. Cada grupo contará con un ayudante asignado para el desarrollo del proyecto. Las fechas y mas detalles sobre el proyecto, se darán a conocer a mediados del semestre, pero la entrega final del proyecto coincide con la fecha apartada para el examen, según la planificación horaria de la Escuela de Ingeniería.

Actividades. En las semanas donde no se publica una tarea, se darán actividades que contribuyen a lograr las competencias mencionadas en los objetivos del curso. Estas actividades son **obligatorias**, y su logro será reportado y clasificado como **L** (logrado), **P** (parcialmente logrado) y **N** (no logrado).

Nota final del curso - sin tarea optativa. Calculemos **AF** como el número resultante de sumar 1 por cada actividad formativa **L**ograda y $0,5$ por cada actividad formativa **P**arcialmente lograda. Llamemos T_i a la nota de la tarea i, y P_i a la nota de la etapa del proyecto (son 3 etapas). Luego la nota final se calcula como

$$\sum_{1\leq i\leq 5} 0{,}14\mathbf{T}_i + 0{,}14\mathbf{AF} - \text{mín}[\mathbf{T}_1, \ldots, \mathbf{T}_5, \mathbf{AF}] + 0{,}03\mathbf{P}_1 + 0{,}1\mathbf{P}_2 + 0{,}17\mathbf{P}_3$$

Nota final del curso - con tarea optativa. En el caso de entregar la tarea optativa, la nota se calcula como el máximo entre la nota sin tarea (de arriba) y:

$$\sum_{1\leq i\leq 6} 7/6\mathbf{T}_i + 7/6\mathbf{AF} - \text{mín}[\mathbf{T}_1, \ldots, \mathbf{T}_6, \mathbf{AF}] + 0{,}03\mathbf{P}_1 + 0{,}1\mathbf{P}_2 + 0{,}17\mathbf{P}_3$$

Calendario Semanal

El material para cada semana se publicará durante los lunes de esa misma semana.

Semana	Contenidos	Evaluación
1	Intro	Actividad
2	Regresión	Actividad, entrega Proyecto1
3	Logit, Descent	Tarea1
4	KNN	Actividad, entrega Tarea1
5	Árboles	Tarea2
6	Bayes	Actividad, entrega Tarea2
7	Clustering	Tarea3
8	GMM	Actividad, entrega Tarea3
9	PCA	Actividad
10	Embeddings	Actividad, entrega Proyecto2
11	GMMs como encoders	Tarea4
12	Texto	Actividad, entrega Tarea4
13	Embeddings - texto 1	Actividad, Tarea5
14	Embeddings - texto 2	entrega Tarea5
15	Kernel Method	Tarea6
16	Grafos	entrega Tarea6
examen		entrega Proyecto3

Todas las entregas son via siding, y por lo general son siempre los Viernes a las 20:00 hrs.

Presencialidad

A la fecha de este escrito, el curso permanece programado como remoto. En caso de pasar a otra modalidad por fuerza mayor, ofrecemos estas garantías:

- Las clases siempre estarán grabadas.
- Velaremos por que todos quienes quieran asistir a clases y/o laboratorios, puedan hacerlo, en la medida que la universidad nos lo permita (aunque puede que tengan que tomar turnos).
- Vamos a implementar un canal de discord para los laboratorios y las ayudantías, el que estará disponible bajo cualquier planificación *excepto* que volvamos a una clase 100 % presencial.

Las clases del curso son obligatorias. En caso de faltar a una clase es responsabilidad del alumno ponerse al día con los contenidos mediante los videos.

Otros

El Departamento de Ciencias de la Computación adopta una política de tolerancia-cero frente a copias o plagios. Se sugiere revisar las políticas y penalidades que el departamento establece ante estas acciones. Recuerda también que la universidad y la escuela están suscritas a un código de honor, lo que nos incluye a profesor, ayudantes y alumnos.
Con respecto a copias y plagios, una reflexión. ¿cuál es la razón por la cuál tomas este curso, en una universidad que cuenta con un grupo de ciencia de datos de nivel mundial? Los ejercicios de este curso están pensados para que puedas ir aprendiendo a medida que te vamos evaluando. Siempre vamos a estar dispuestos a contestar todas tus dudas. ¡Aprovecha esta oportunidad para aprender!

El curso tiene dos canales de comunicación oficiales: Las clases y la página Web. Se asume que que toda la información que es entregada por ambos canales llega a todos los alumnos. Por lo mismo, se sugiere a los alumnos revisar la página Web constantemente. Si bien usamos discord, no lo utilizaremos como un medio para propagar información importante.

Bibliografía

1. Aggarwal, C. C. (2015). Data mining: the textbook. Springer.
2. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.
3. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Second Edition. New York: Springer series in statistics, 2009.