IIC2433 — Minería de Datos

Programa de Curso

Agosto 2024

Página de Curso

Importante: la página del curso está en `https://github.com/IIC2433/Syllabus-2024-2`. 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 prácticas 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 reu ne dos veces a la semana. La clase es los d ıas martes, de 14:00 a 16:50, y los jueves de 14:00 a 15:20 habr a un horario de trabajo y consulta. La clase, en si misma, se divide en dos bloques, uno para ver contenidos te oricos y el otro para realizar un laboratorio pr actico. Es decir, semana a semana, las alumnas tendr an el siguiente esquema:
El curso se reúne una vez a la semana, los martes de 14:50 a 17:20. La clase se divide en dos bloques, uno para realizar contenidos teóricos y otro para un laboratorio práctico. Es decir, semana a semana, el esquema es el siguiente:

- Martes 14:50: Clase expositiva.
- Martes después de la clase expositiva: trabajo práctico.
- Viernes: entrega del trabajo práctico.

Ayudantías. Adicionalmente, en el horario Jueves 14:50, realizaremos algunas instancias de consulta y de discusión de tareas. Estas serán avisadas oportunamente, cuando no haya aviso, no hay clases los Jueves.

Evaluación

El curso se evalúa a través de:

- Cinco tareas individuales.
- Trabajo formativo semana a semana.
- Una proyecto final en grupos de a 4 personas.

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 (4 personas) 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. 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 $\mathbf{T_i}$ a la nota de la tarea i, y $\mathbf{P_i}$ a la nota de la etapa i del proyecto (son 3 etapas). Luego la nota final se calcula como

$$0{,}7\Big(\sum_{1\leq i\leq 5} T_i + AF - \text{mín}[AF, T_1, \ldots, T_5]\Big) + 0{,}03P_1 + 0{,}1P_2 + 0{,}17P3$$

Calendario del curso

Importante: Este calendario puede sufrir modificaciones, las que se avisarán oportunamente.
El calendario se estructura en base a semanas:

Semana de clase	Tema	Evaluación
Semana 1	Introducción, Pandas	Actividad
Semana 2	Regresión	Actividad
Semana 3	Logit, Gradient Descent	Tarea 1, Entrega Proyecto 1
Semana 4	KNN	Actividad, entrega Tarea 1
Semana 5	Árboles	Tarea 2
Semana 6	Explicar, inferencia causal	Actividad, entrega Tarea 2
Semana 7	Clustering 1	Tarea 3
Semana 8	Clustering 2 (en clases)	Actividad, Entrega Tarea 3
Semana 9	Reducción de dimensionalidad 1	Actividad
Semana 10	Reducción de dimensionalidad 2	Actividad, Entrega Proyecto 2
Semana 10	Espacio Latente	Tarea 4
Semana 12	Embeddings texto 1	Actividad, entrega Tarea 4
Semana 13	Embeddings texto 2	Tarea 5
Semana 14	Otros Embeddings	Actividad, Entrega Tarea 5
Semana 15	Métodos de Kernel	Actividad
Semana 16	Otros tópicos	

Contenidos

1. **Modelos de predicción, clasificación, explicación**
   a) Regresión lineal y logística
   b) Métodos de afinamiento y testeo
   c) Vecinos más cercanos
   d) Árboles de decisión
   e) Inferencia causal
2. **Clustering**
   a) K-means, DBSCAN
   b) Modelos de mixturas de distribuciones
   c) evaluación
3. **Preparación y transformación de la información**
   a) Librerías para el trabajo con datos
   b) Análisis de componentes principales
   c) Autoencoders
   d) Métodos de Kernel
4. **Información semi-estructurada**
   a) Texto
   b) Grafos
5. **Tópicos Avanzados**
   a) Por anunciar, de acuerdo a disponibilidad de semestre.

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 investigación en 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.

Las clases del curso son obligatorias. En caso de faltar a una clase es responsabilidad del alumno ponerse al día con los contenidos. No se borran evaluaciones, pero se aceptará rendir evaluaciones de forma atrasada, cuando la situación lo amerite.

Bibliografía mínima

1. Rajaraman, A, and Ullman, J.D. Mining of massive datasets. Cambridge University Press, 2011. (disponible online).
2. Aggarwal, C. Data mining: the textbook. Springer, 2015 (en la biblioteca).
3. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. The elements of statistical learning: data mining, inference, and prediction Springer, 2009 (en la biblioteca).