REGISTRADO BAJO Nº CDCIC-145/20

BAHIA BLANCA, 20 de agosto de 2020

VISTO:

El informe de evaluación elaborado por el CONEAU (Consejo Nacional de Evaluación y Acreditación Universitaria), con expediente N° EX-2019-98114342-APN-DAC#CONEAU, referido a la solicitud de acreditación de la Especialización en Ciencias de Datos;

La nota elevada por el Secretario de Investigación y Posgrado informando los cambios que el Comité Académico de Dirección de la Especialización en Ciencia de Datos ha propuesto para los cursos de la misma;

CONSIDERANDO:

Que dicho Comité recomendó estos cambios teniendo en cuenta las observaciones realizadas por los pares evaluadores de CONEAU;

Que es atribución del Consejo Departamental de Ciencias e Ingeniería de la Computación aprobar las pautas que regulen el adecuado funcionamiento de las carreras de posgrado dictadas por esta Unidad Académica;

Que el Consejo Departamental aprobó en su reunión de fecha 20 de agosto de 2020 lo propuesto;

POR ELLO,

EL CONSEJO DEPARTAMENTAL DE
CIENCIAS E INGENIERÍA DE LA COMPUTACIÓN

R E S U E L V E:

Art. 1°).- Avalar los cambios que el Comité Académico de Dirección de la Especialización en Ciencia de Datos ha propuesto para los cursos de dicha carrera, de acuerdo Anexo I que se adjunta.-

Art. 2º).- Regístrese y pase a la Secretaría General de Posgrado y Educación Continua de la UNS a los fines que corresponda. Cumplido, archívese.----------

Firmado digitalmente por MARTÍNEZ Diego César
Nombre de reconocimiento (DN): serialNumber=CUIL 23246916829, c=AR, cn=MARTÍNEZ Diego César

Dr. Marcelo A. Falappa
Director Decano
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur

Digitally signed by FALAPPA Marcelo Alejandro

ANEXO

5.1 Matemática aplicada e introducción a la ciencia de datos

Este primer curso presenta contenidos matemáticos sobre los cuales se sustentan los métodos de ciencia de datos. La idea de este primer curso no es la revisión completa y minuciosa de los temas que aquí se presentan, sino una presentación de los conceptos más importantes enfocándose en destacar la intuición del formalismo y su aplicación, más que la justificación matemática. El objetivo es que los alumnos puedan acceder a los contenidos mínimos y necesarios para comprender métodos más complejos usados en la ciencia de datos.

El álgebra lineal es una disciplina matemática fundamental para la representación de datos, así como la búsqueda y optimización dentro del espacio representado matemáticamente. Por este motivo comenzaremos este curso con el estudio de técnicas de álgebra lineal aplicadas a problemas recurrentes en ciencia de datos. Por otra parte, la ciencia de datos se basa fuertemente en el uso apropiado de métodos de estadística. Las implicancias de los resultados alcanzados dependen de la comprensión y alcance de los fundamentos estadísticos aplicados. Finalmente, los métodos de análisis estadístico y aprendizaje automático requieren de un alto grado de computación numérica. Dada la capacidad finita de almacenamiento de una computadora, las operaciones que involucran números reales pueden dar lugar a problemas de precisión, y por lo tanto, es importante reconocer y saber cómo remediar esta situación.

Objetivos

Con este curso se espera que el alumno logre:

- representar datos utilizando espacios vectoriales n-dimensionales;
- analizar y manipular datos utilizando herramientas de álgebra lineal;
- identificar fuentes de incertidumbre en los datos;
- modelar datos usando funciones de distribución de probabilidad;
- conocer métricas estadísticas, así como sus supuestos y aplicaciones;
- conocer los tests de hipótesis que permitan identificar diferencias significativas en los resultados;
- medir distancias entre representaciones de incertidumbre;
- conocer las bases de una optimización matemática;
- reconocer situaciones de pérdida de precisión y cómo pueden ser remediadas.

Contenidos Mínimos

- **Álgebra lineal aplicada.** Creación de matrices y vectores a partir de atributos obtenidos de conjuntos de datos. Cálculo de métricas de similitud entre representaciones vectoriales. Determinación de dependencia lineal. Transformaciones de matrices: escalado, dilatación y cambio de coordenadas. Selectores: reducción de la muestra, recorte y permutaciones. Reducción de dimensionalidad. Descomposiciones de matrices: descomposición en valores singulares, análisis de componentes principales y factorización de matrices no negativas. Métodos de ajuste basados en mínimos cuadrados. Independencia probabilística. Regla de Bayes.

- **Estadística aplicada.** Métodos de muestreo. Inferencia estadística. Estimación estadística de parámetros. Intervalo de confianza. Testeo de hipótesis. Análisis de asociación entre variables.
- **Métodos numéricos.** *Overflow y underflow*. Número de condición. Optimización basada en gradiente. Optimización de segundo orden. Optimización con restricciones.

Actividades

Esta primera materia sirve para presentar los conceptos de álgebra lineal, estadística y métodos numéricos necesarios para abordar problemas de ciencia de datos. Los conceptos serán abordados desde un enfoque práctico. Ciertas áreas de la ingeniería y las ciencias exactas se abocan en actividades de índole determinista y sin incertidumbre. Por el contrario, la ciencia de datos requiere de la familiarización con situaciones inciertas y no deterministas. Por lo tanto, es importante familiarizar al alumno con sistemas inherentemente estocásticos, situaciones con observabilidad incompleta, y modelos aproximados.

Desde el punto de vista teórico, se omitirán conceptos que no sean estrictamente necesarios para los contenidos de la especialización. Para esto se trabajará a nivel de intuición más que en el desarrollo matemático completo, aunque los docentes brindarán las herramientas y literatura adicional para que el alumno pueda consultar y profundizar más en el tema de así necesitarlo.

Las actividades prácticas se llevarán a cabo en un laboratorio de computación y estarán orientadas a proponer y desarrollar ejemplos utilizando el lenguaje de programación Python. En este sentido, se mostrará cómo los formalismos matemáticos estudiados sirven para modelar y manipular distintos tipos de datos. Para esto se plantearán conjuntos de datos modelo para que el alumno utilice bibliotecas disponibles para el cómputo científico, aplique las técnicas aprendidas y reflexione sobre los resultados. Para facilitar la presentación documentada de ejemplos y la generación de soluciones por parte de los alumnos, utilizaremos la plataforma Google Colab, la que se basa en el uso de cuadernos Jupyter. Mediante Google Colab los alumnos podrán acceder a una serie de bibliotecas que serán utilizadas para el modelado, transformación, análisis y visualización de datos. Entre las bibliotecas que esperamos utilizar en este curso cabe mencionar NumPy y SciPy.

Para la modalidad no presencial, se plantearán ejercicios de programación que extenderán o complementarán aquellos vistos durante las clases prácticas presenciales. Para el desarrollo de los ejercicios en modalidad no presencial se utilizarán las mismas herramientas y bibliotecas utilizadas durante las clases presenciales. Además de la interacción que se dará naturalmente al compartir los cuadernos de Google Colab, se utilizarán los foros y mensajes de Moodle para que los alumnos puedan realizar preguntas y para que los docentes puedan despejar dudas y brindar la guía necesaria para resolver los ejercicios planteados. Se llevarán a cabo videoconferencias para eventuales consultas de mayor profundidad.

Bibliografía

- Agresti, A., Franklin, C. & Klingenberg, B. (2018). Statistics the Art and Science of Learning from data. Pearson Education Limited.

- Boyd, S., & Vandenberghe, L. (2018). Introduction to applied linear algebra: vectors, matrices, and least squares. Cambridge university press.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. Part 1
- Petersen, K. B., & Pedersen, M. S. (2012). The Matrix Cookbook.
- Shilov, G. E. (2012). Linear Algebra. Dover Publications.

Cabe destacar que toda la bibliografía indispensable para la realización del curso estará disponible online o en la Biblioteca del DCIC.

5.2 Adquisición y limpieza de datos

Muchos enfoques y libros de texto abocados al análisis de datos parten de la suposición errónea de que los datos están disponibles y listos para su uso. En realidad, antes de poder aplicar un método de análisis, se requiere, en primer lugar, de la obtención, limpieza y la conversión a un formato conveniente para su análisis. Distintas encuestas[1] reportaron que los científicos de datos pasan el 80% de su tiempo buscando, limpiando y organizando datos, mientras que el 20% restante es usado para el análisis en sí. Una encuesta más reciente y aún más masiva[2] sitúa a los "datos sucios" como el principal problema con el que los científicos de datos tienen que lidiar (49,4%), mientras que la dificultad al acceso de los datos se sitúa en el cuarto lugar (30,4%) de una lista de 15 problemas principales.

En este curso se espera que el alumno conozca distintas fuentes de acceso de datos, así como también identifique problemas en los mismos y las herramientas necesarias para su depuración, limpieza y organización. Los alumnos también tendrán la opción de modificar y ejecutar algoritmos de limpieza según casos recreados a partir de datos reales.

Objetivos

Con este curso se espera que el alumno logre:

- conocer y comprender el origen de dificultades para el análisis de los datos;
- saber cómo compartir los datos para uso interno o colaboración académica;
- entender las características principales de distintos formatos de datos;
- conocer los principales repositorios de datos públicos;
- reconocer posibles problemas en los datos y posibles métodos de limpieza e integración;
- saber amortiguar el impacto de datos sucios o incompletos en los tiempos de proyecto;
- ejercitar la aplicación de algoritmos sencillos usando librerías de adquisición y limpieza de datos.

[1] https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

[2] https://www.kaggle.com/surveys/2017?utm=cade

Contenidos Mínimos

- **Fuentes de datos**. Internos/externos, públicos/privados. Bases de datos, internet, rascado de sitios web, APIs (*application programming interfaces*), encuestas. Ejemplos de dominio público: API de redes sociales, datos de gobierno abierto, competencias académicas (e.g. TREC, SemEval) y comerciales (e.g. Kaggle).
- **Formatos de datos**. Planillas de cálculo, archivos de texto, XML, JSON. Estructuras de datos específicas para ciencia de datos.
- **Compartimiento de datos**. Datos crudos, datos procesados, documentación y recetario. Repositorios públicos de datos.
- **Carga de datos**. Carga de datos en memoria en el contexto de Big Data. Codificaciones de texto.
- **Datos segmentados.** Fusión de datos.
- **Operaciones básicas**. Obtención de subconjunto. Concatenación. Ordenamiento. Completando datos incompletos. Uniformidad de datos. Alineamiento de datos. Normalización de fechas. Expresiones regulares.

Actividades

El contenido del curso brindará conocimientos teóricos y prácticos en el contexto de las distintas fuentes, formatos y problemas que pueden presentar los datos a analizar. Los temas teóricos serán dictados por el docente a cargo, en un laboratorio de computación, e incluirán un resumen de los principales problemas que los científicos de datos enfrentan a la hora de analizar un conjunto de datos. Los mismos serán acompañados de ejemplos que ayuden a comprender la naturaleza del problema. Los temas prácticos incluirán el desarrollo de guías de trabajos que apunten a desarrollar capacidades para la extracción e integración de datos provenientes de distintas fuentes, filtrado de datos, manejo de datos incompletos, identificación de datos redundantes, uniformización de valores y la organización de los datos en formatos o estructuras de datos específicas. Las mismas serán realizadas utilizando scripts en python (Jupyter Notebooks), utilizando conjuntos de datos de casos reales y apoyándose en distintas librerías para tal fin (e.g., numpy, scipy, pandas, re, datetime). Para este propósito se utilizarán librerías de manejo de datos que facilitan la manipulación y limpieza de grandes cantidades de datos. Para poder aplicar dichos algoritmos, los alumnos recibirán una instrucción de tipo tutorial-práctica que les permitirá comprender y aplicar algoritmos sencillos de limpieza y organización de datos.

Las actividades prácticas se realizarán tanto en forma presencial en un laboratorio de computación como a distancia mediante acceso remoto usando Google Colab u otras herramientas. La presentación inicial de las actividades y problemas a realizar a distancia, como la posterior discusión y puesta en común de los resultados, se realizará durante las clases prácticas presenciales. Además, los docentes supervisarán a los alumnos mediante foros de Moodle, mensajes directos y a través de los cuadernos de Google Colab compartidos.

Se detallan a continuación las principales actividades prácticas, indicando también si la misma es presencial o no:

A) Uso del software Postman y scripts en Python para el manejo de RESTful APIs. Actividad presencial.

B) Tutorial uso de librería Pandas para el manejo de datos incompletos, remoción de variables y/o columnas según condiciones, estandarización numérica y normalización de texto. Actividad presencial.

C) Exploración de datos y visualizaciones básicas para identificar anormalidades, datos faltantes o falta de uniformidad en los datos recolectados. Actividad presencial.

D) El alumno implementará la solución a un problema simple pero que requiera integrar las herramientas de software y técnicas presentadas anteriormente. Esta tarea involucra no sólo la parte técnica sino también la presentación por informe escrito de los problemas identificados y las estrategias de resolución utilizadas. Actividad no presencial.

E) Puesta en común y discusión de la solución a los problemas planteados y los resultados obtenidos. Actividad presencial.

Las actividades presenciales se realizarán en los laboratorios del DCIC, mientras que las actividades no presenciales se coordinarán a través de los foros de Moodle, mensajes privados con los docentes y consultas por videoconferencia. La evaluación y el seguimiento de estas actividades se realizará a través del mismo sistema, usando los foros, cuestionarios y entregas de los resultados obtenidos.

Bibliografía

- Lutz, M. (2013). Learning Python (5th Edition). O'Reilly Media.
- McKinney, W. (2012). Python for Data Analysis. O'Reilly Media.
- Peersman, G. (2014). Overview: Data Collection and Analysis Methods in Impact Evaluation: Methodological Briefs - Impact Evaluation No. 10. UNICEF Office of Research-Innocenti.
- VanderPlas, J. (2016). Python Data Science Handbook. O'Reilly Media.
- Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005). Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLoS Medicine, 2(10), e267.

Cabe destacar que toda la bibliografía indispensable para la realización del curso estará disponible online o en la Biblioteca del DCIC.

5.3 Almacenamiento de Datos para Big Data

A medida que los repositorios de datos se vuelven más voluminosos y complejos, su almacenamiento en archivos de texto se torna inapropiado. En este contexto, las bases de datos representan la opción más natural. Hay muchos tipos de sistemas de manejo de base datos, pero en los últimos años se empezó a trazar una división entre las bases de datos tradicionales (también llamadas relacionales) y las llamadas NoSQL. El surgimiento de las de tipo NoSQL se correlaciona con el avance en la complejidad de los datos y la necesidad de analizarlo.

Los contenidos de este curso introducen al alumno a los distintos paradigmas para el almacenamiento de datos en el contexto de Big Data. En particular, se enfatiza la necesidad de elegir el tipo de almacenamiento apropiado de acuerdo a la tarea o a la naturaleza de los datos. También

se busca desarrollar conocimientos aplicados en el uso de paradigmas diferentes (e.g. tipo documento o tipo grafo) dada su utilización actual y la versatilidad de las mismas.

Objetivos

Con este curso se espera que el alumno logre:

- Entender conceptos relacionados al almacenamiento persistente en base de datos.
- Reconocer las ventajas y limitaciones de utilizar una base de datos de tipo NoSQL en el contexto de Big Data.
- Identificar el mejor tipo o paradigma de base de datos NoSQL para la naturaleza de los datos y del análisis.
- Desarrollar habilidades prácticas para la escritura y lectura en distintas base de datos de uso actual y masivo.

Contenidos Mínimos

- **Bases de datos (BD)**. Limitaciones en el almacenamiento en archivos de texto. BD relacionales. Sintaxis SQL.
- **BD NoSQL**. Historia. Comparación con BD relacionales: compartimiento de datos, propiedades ACID/BASE, escalabilidad, flexibilidad, adaptación al cambio y entornos ágiles. Mitos y verdades sobre las BD NoSQL. Paradigmas: llave/valor, documento, columna, grafos, otras.
- **BD de tipo documento (e.g. MongoDB)**. Shell, drivers, replicación, sharding. Documentos y colecciones. Inserción, modificación y recuperación de documentos. Distintos tipos de indización.
- **BD de tipo grafo (e.g. Neo4J)**. Drivers y arquitectura. Inserción de datos y consulta. Visualización de Grafos.

Actividades

Las actividades docentes de este curso apuntan a brindar una introducción al mundo de las bases de datos en el contexto de Big Data y ciencia de datos. Por tal motivo, las clases teóricas, a ser dictadas en un laboratorio de computación, se centrarán en la explicación de las principales diferencias entre distintos tipos de base de datos, en especial las de tipo NoSQL. Para esto se incluirán comparaciones que permiten contrastar los objetivos que persigue cada tipo de base de datos. Dichos objetivos servirán de base para que el alumno logre la capacidad de discernir e identificar el tipo más propicio de base de datos para su problema bajo análisis.

Los temas prácticos brindarán al alumno la posibilidad de incursionar en la inserción, modificación y recuperación de datos en bases de datos NoSQL utilizando datos reales. Se prevé que dicha práctica se realice parte en forma presencial y parte en forma a distancia asincrónica. Las actividades prácticas se centrarán en las bases de datos de tipo documento y de tipo grafo. Para las de tipo documento se utilizará MongoDB, el cual permite el uso de un enfoque flexible para almacenar datos anidados y de distintos tipos, incluyendo texto. Para las de tipo grafo se utilizará Neo4j, el cual brinda un enfoque inherentemente relacional y permite desagregar la información y modelar datos interconectados.

Las actividades prácticas se realizarán tanto en forma presencial en un laboratorio de computación como a distancia mediante acceso remoto usando Google Colab y accediendo a un servidor del DCIC con bases de datos NoSQL ya instaladas. Las actividades apuntarán a desarrollar habilidades prácticas para el almacenamiento y recuperación efectiva de datos grandes, multivariados e interconectados. La presentación inicial de las actividades y problemas a realizar a distancia, como la posterior discusión y puesta en común de los resultados, se realizará durante las clases prácticas presenciales. Además, los docentes supervisarán a los alumnos mediante foros de Moodle y mensajes directos.

Se detallan a continuación actividades prácticas con los tiempos y modalidad especificados:

A) A partir del planteo de distintos casos de uso donde se propongan tareas de análisis de datos, el alumno deberá elegir la o las estrategias de almacenamiento más propicias. Actividad presencial.

B) Se trabajará en clase en un taller tutorial en donde se brinde enseñanza práctica para el almacenamiento y recuperación de bases de datos NoSQL de distintos tipos: documento, clave-valor, grafos y columna. Actividad presencial.

C) El alumno implementará la solución a un problema que requiera el almacenamiento de los datos en una base de datos de tipo documento, así como la provisión de rutinas de recuperación de datos. Esta tarea involucra no sólo la parte técnica sino también la presentación por informe escrito de los problemas identificados y las estrategias de resolución utilizadas. Actividad no presencial.

D) El alumno implementará la solución a un problema que requiera el almacenamiento de los datos en una base de datos de tipo grafo, así como la provisión de rutinas de recuperación de datos. Esta tarea involucra no sólo la parte técnica sino también la presentación por informe escrito de los problemas identificados y las estrategias de resolución utilizadas. Actividad no presencial.

E) Puesta en común y discusión de soluciones a los problemas planteados y los resultados obtenidos. Actividad presencial.

La evaluación y el seguimiento de estas actividades se realizará a través del mismo sistema, usando los foros, cuestionarios y entregas de los resultados obtenidos.

Bibliografía

- Atzeni, P., Jensen, C. S., Orsi, G., Ram, S., Tanca, L., & Torlone, R. (2013). The relational model is dead, SQL is dead, and I don't feel so good myself. ACM SIGMOD Record, 42(1), 64.
- Baton, J., & Van Bruggen, R. (2017). Learning Neo4j 3.x: effective data modeling, performance tuning and data visualization techniques in Neo4j. Packt.
- Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., … Gruber, R. E. (2008). Bigtable: A Distributed Storage System for Structured Data. ACM Transactions on Computer Systems, 26(2).
- Chodorow, K., & Bradshaw, S. (2018). MongoDB: The Definitive Guide (3rd Edition). O'Reilly.
- Floratou, A., Teletia, N., DeWitt, D. J., Patel, J. M., & Zhang, D. (2012). Can the elephants handle the NoSQL onslaught? Proceedings of the VLDB Endowment, 5(12), 1712–1723.
- MongoDB. (2015). Top 5 Considerations When Evaluating NoSQL Databases.
- Robinson, I., Webber, J., & Eifrem, E. (2015). Graph Databases. O'Reilly Media.
- Sadalage, P. J., & Fowler, M. (2013). NoSQL distilled: a brief guide to the emerging world of polyglot persistence. Addison-Wesley.
- Strauch, C., & Kriha, W. (2009). NoSQL Databases Lecture Selected Topics on Software-

Technology Ultra-Large Scale Sites.

Cabe destacar que toda la bibliografía indispensable para la realización del curso estará disponible online o en la Biblioteca del DCIC.

5.4 Aprendizaje automático

Este curso presenta los conceptos fundamentales, técnicas metodologías y herramientas de aprendizaje automático. El aprendizaje automático provee la base técnica que permite extraer información útil de los datos. Se estudiarán conceptos y métodos de aprendizaje no supervisado (reducción de la dimensionalidad, métodos de kernel y métodos de clustering) y supervisado (clasificadores, fusión, meta-aprendizaje) que permitirán encontrar patrones estructurales en datos. El curso también presentará una introducción a conceptos y técnicas de redes neuronales artificiales, aprendizaje por refuerzo y minería de datos. Por otra parte, se introducirá al alumno a las metodologías empleadas para construir y validar modelos de aprendizaje automático. El curso a la vez entrenará al alumno en el uso de herramientas actuales que facilitan la aplicación de las técnicas estudiadas.

Además del estudio de conceptos, técnicas y metodologías de aprendizaje automática y minería de datos, el alumno realizará trabajos prácticos orientados a ganar experiencia y entrenamiento en el uso de las tecnologías presentadas en aplicaciones reales.

Objetivos

Con este curso se espera que el alumno logre:

- adquirir conceptos de aprendizaje automático y minería de datos;
- decidir qué métodos de aprendizaje automático son apropiados para diferentes tipos de problemas de aprendizaje;
- conocer las limitaciones y desventajas de los diferentes métodos de aprendizaje automático;
- decidir cómo representar los datos para facilitar el aprendizaje;
- identificar soluciones para tratar problemas tales como el ruido, los outliers y los datos de alta dimensionalidad;
- construir modelos de aprendizaje automático;
- validar modelos de aprendizaje automático;
- reconocer efectos típicos ocasionados por sesgos en las inicializaciones y selección de parámetros;
- sugerir refinamientos de los modelos de aprendizaje automático para mejorar los resultados observados en el proceso de validación;
- adquirir experiencia práctica en el uso de herramientas de aprendizaje automático.

Contenidos Mínimos

- **Conceptos y técnicas de aprendizaje no supervisado.** Reducción de la dimensionalidad. Métodos de kernel. Clustering aglomerativo. Clustering divisivo. Clustering por re-locación (k-means, EM). Clustering por densidad (DBScan). Clustering via embeddings (SOMs, LSI).
- **Conceptos y técnicas de aprendizaje supervisado**. Clasificadores basados en similitud (k-

vecinos más cercanos). Clasificadores probabilísticos (bayesianos). Clasificadores lineales (perceptrón, SVM). Clasificadores lineales generalizados (métodos de kernel). Clasificadores basados en teoría de la información (árboles de decisión). Fusión y meta-aprendizaje (stacking, bagging, boosting, ECOC). Análisis de regresión. Regresión logística.

- **Redes neuronales artificiales.** Conceptos. Arquitecturas. Redes profundas. Técnicas de entrenamiento. Validación.
- **Aprendizaje por refuerzo.** Mecanismos de exploración**.** Algoritmos para controlar el aprendizaje.
- **Minería de datos.** Itemsets frecuentes. Minería de reglas de asociación.
- **Validación.** Cómputo de métricas de performance. Estrategias de validación estadística.
- **Herramientas de aprendizaje automático.** Ejemplos sugeridos: Orange, Scikit-Learn, Keras, Pyqlearning y Mlxtend.

Actividades

El dictado de clases teóricas se complementa con clases prácticas que se llevarán a cabo en un laboratorio de computación. Como parte de las actividades prácticas se propondrán y desarrollarán ejemplos utilizando herramientas de aprendizaje automático y el lenguaje de programación Python. Se utilizarán herramientas con ambientes visuales amigables, tales como Orange, para una primera aproximación al estudio de las técnicas de aprendizaje automático estudiadas. Luego se pasará al desarrollo de ejercicios que se apoyarán en el uso de bibliotecas tales como Scikit-Learn, Pyqlearning y Mlxtend. Estos ejercicios requerirán la implementación de soluciones de aprendizaje automático mediante la programación en el lenguaje de programación Python.

Se utilizarán conjuntos de datos etiquetados y no etiquetados para desarrollar modelos de aprendizaje automático y validarlos. Para facilitar la presentación documentada de ejemplos y la generación de soluciones por parte de los alumnos, utilizaremos la plataforma Google Colab, la que se basa en el uso de cuadernos Jupyter. Mediante Google Colab los alumnos podrán acceder a bibliotecas, tales como Scikit-Learn, Pyqlearning y Mlxtend, que serán utilizadas para experimentar con los algoritmos estudiados.

Para la modalidad no presencial, se plantearán ejercicios de programación que extenderán o complementarán aquellos vistos durante las clases prácticas presenciales. Para el desarrollo de los ejercicios en modalidad no presencial se utilizarán las mismas herramientas y bibliotecas utilizadas durante las clases presenciales. Además de la interacción que se dará al compartir los cuadernos de Google Colab, se utilizarán los foros y mensajes de Moodle para que los alumnos puedan realizar preguntas y para que los docentes puedan despejar dudas y brindar la guía necesaria para resolver los ejercicios planteados. Se llevarán a cabo videoconferencias para eventuales consultas de mayor profundidad.

Bibliografía

- Baeza-Yates, R., & Ribeiro-Neto, B. (2011). Modern information retrieval: the concepts and technology behind search. Addison Wesley.
- Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
- Büttcher, S., Clarke, C. L. A., & Cormack, G. V. (2010). Information retrieval: implementing

and evaluating search engines. MIT Press.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
- Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of Massive Datasets (2nd ed.). Cambridge University Press.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
- McKinney, W. (2012). Python for Data Analysis. O'Reilly Media.
- Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.
- VanderPlas, J. (2016). Python Data Science Handbook. O'Reilly Media.
- Witten, I. H., Frank, E., Hall, M. A. , & Pal, C. J. (2016). Data mining: practical machine learning tools and techniques. Morgan Kaufmann.

Cabe destacar que toda la bibliografía indispensable para la realización del curso estará disponible online o en la Biblioteca del DCIC.

5.5 Análisis de datos no estructurados y no convencionales

El curso "Análisis de datos no estructurados y no convencionales" introduce conceptos y técnicas relevantes para la manipulación, exploración y análisis de datos que se presentan en forma de texto, audio, imágenes y video, como así también de datos interconectados en forma de red y series de tiempo. Se estudiarán técnicas orientadas a representación y análisis de texto y recuperación de información. Por otra parte, se estudiarán técnicas que permitirán procesar, extraer atributos útiles y construir modelos predictivos a partir de datos multimediales (audio, imágenes y video). Se presentarán los conceptos básicos del área de redes complejas y se examinarán datos interconectados con el objeto de adquirir experiencia en el análisis de datos en forma de red, utilizando redes sociales, tecnológicas y biológicas. Finalmente, se examinarán técnicas para el análisis de series de tiempo. Para complementar el entrenamiento práctico del alumno se aplicarán algunas de las técnicas vistas en el curso "Aprendizaje automático" a los datos no estructurados y no convencionales que son específicos de este curso.

Además del estudio de conceptos y técnicas para el análisis de datos no estructurados y no convencionales el alumno realizará trabajos prácticos orientados a ganar experiencia en la aplicación de las tecnologías presentadas sobre conjuntos de datos reales.

Objetivos

Con este curso se espera que el alumno logre:

- adquirir conceptos de minería de texto y recuperación de información;
- decidir qué técnicas de procesamiento son apropiadas para diferentes tipos de datos multimediales;
- decidir cómo representar los datos multimediales para facilitar la construcción de modelos y la predicción;
- identificar soluciones para sobreponerse a la alta dimensionalidad de los datos multimediales;
- adquirir experiencia práctica en el uso de herramientas de predicción a partir de datos multimediales;
- identificar características salientes de datos en forma de red;

- conocer las métricas y algoritmos que permiten identificar propiedades locales y globales en datos en forma de red;
- familiarizarse con herramientas que permitan explorar datos en forma de red;
- adquirir conceptos y aprender técnicas para el análisis de series de tiempo.

Contenidos Mínimos

- **Texto**. Representación de texto. Conceptos de recuperación de información. Índices invertidos. Recuperación booleana. Modelo vectorial. Modelo probabilístico. Evaluación. Herramientas para implementación de buscadores y exploración visual del contenido indexado (e.g. Elasticsearch/Kibana). Análisis de sentimiento. Categorización y clustering. Reconocimiento de entidades. Modelamiento de temas. Extracción de conceptos (e.g. wikificación). Técnicas avanzadas (e.g. QA, generación automática de resúmenes). Herramientas para análisis de texto (e.g. GATE o RapidMiner).
- **Audio, Imágenes y video**. Procesamiento, extracción de atributos y predicción con datos multimediales.
- **Datos interconectados.** Conceptos básicos de redes complejas (propiedades, métricas y algoritmos). Minería de redes sociales, tecnológicas y biológicas. Análisis de enlaces. Detección de comunidades y nodos centrales. Herramientas (e.g. Gephi)
- **Análisis de series de tiempo.** Tendencias, efectos estacionales, cíclicos y residuales. Modelos aditivos y multiplicativos. Modelo autorregresivo de media móvil.

Actividades

El dictado de clases teóricas buscará impartir los conocimientos necesarios para comprender tanto técnicas clásicas como las más actuales para el análisis de datos no estructurados y no convencionales. El aprendizaje práctico de estas técnicas se logrará mediante actividades presenciales y a distancia, utilizando datos de distinta naturaleza.

Para adquirir conocimiento práctico en técnicas de recuperación de información se utilizarán herramientas tales como Elasticsearch y Kibana para la construcción de buscadores sobre documentos de texto y para la exploración visual y el análisis del contenido indexado. Por otra parte, se diseñarán tareas orientadas a convertir texto no estructurado en datos significativos mediante las técnicas de minería de texto estudiadas. Para esto último se utilizarán herramientas tales como GATE o RapidMiner.

Los alumnos experimentarán con datos de audio, imágenes y video con el fin de obtener atributos significativos que permitirán construir representaciones, las que serán puestas a prueba en tareas de predicción. Las actividades que involucren el uso de datos interconectados se llevarán a cabo utilizando la herramienta Gephi (o similar) para la visualización y exploración de grafos, como así también para el cálculo de métricas e identificación de propiedades sobre los datos. Finalmente, se estudiarán las diferentes componentes de una serie temporal utilizando herramientas estadísticas y gráficas empleando datos orientados a aplicaciones reales.

El dictado de clases prácticas se llevará a cabo en un laboratorio de computación. Para la modalidad no presencial, se plantearán ejercicios de programación que extenderán o complementarán aquellos vistos durante las clases prácticas presenciales. Para el desarrollo de los ejercicios en modalidad no presencial se utilizarán las mismas herramientas y bibliotecas utilizadas

durante las clases presenciales. Además de la interacción que se dará naturalmente al compartir los cuadernos de Google Colab, se utilizarán los foros y mensajes de Moodle para que los alumnos puedan realizar preguntas y para que los docentes puedan despejar dudas y brindar la guía necesaria para resolver los ejercicios planteados. Se llevarán a cabo videoconferencias para eventuales consultas de mayor profundidad.

Bibliografía

- Baeza-Yates, R., & Ribeiro-Neto, B. (2011). Modern information retrieval: the concepts and technology behind search. Addison Wesley.
- Bontempi, G., Ben Taieb, S., & Le Borgne, Y.-A. (2013). Machine Learning Strategies for Time Series Forecasting (pp. 62–77). Springer, Berlin, Heidelberg.
- Beyeler, M. (2017). Machine Learning for OpenCV. Packt Publishing.
- Büttcher, S., Clarke, C. L. A., & Cormack, G. V. (2010). Information retrieval: implementing and evaluating search engines. MIT Press.
- Chakrabarti, S. (2003). Mining the Web: discovering knowledge from hypertext data. Morgan Kaufmann.
- Croft, B., Metzler, D., & Strohman, T. (2010). Search Engines Information Retrieval in Practice. Addison Wesley.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
- Hamilton, J. D. (1994). Time Series Analysis. Princeton University Press.
- Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of Massive Datasets (2nd ed.). Cambridge University Press.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
- Prince, Simon, J. D. (2012). Computer Vision: Models, Learning, and Inference. Cambridge University Press.
- Roth, P. M., & Bischof, H. (2008). Conservative Learning for Object Detectors. In Machine Learning Techniques for Multimedia (pp. 139–158). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Valenti, R., Sebe, N., Gevers, T., & Cohen, I. (2008). Machine Learning Techniques for Face Analysis. In Machine Learning Techniques for Multimedia (pp. 159–187). Berlin, Heidelberg: Springer Berlin Heidelberg.
- VanderPlas, J. (2016). Python Data Science Handbook. O'Reilly Media.

Cabe destacar que toda la bibliografía indispensable para la realización del curso estará disponible online o en la Biblioteca del DCIC.

5.6 Exploración, Simulación y Recomendación

Este curso está orientado a estudiar conceptos, prácticas y herramientas para la exploración visual de datos, el modelado y simulación de sistemas y los sistemas de recomendación. Los contenidos de este curso se apoyan fuertemente en contenidos anteriores y representan un nivel más avanzado para el análisis de datos y resolución de problemas complejos. Por dichos motivos, los conceptos y herramientas comprendidos en este curso resultan de interés práctico dentro del ámbito industrial y gubernamental.

El curso está dividido en tres partes. En la primera se cubren conceptos de visualización de datos y minería de datos interactiva. Esto representa un cambio de paradigma para el análisis de problemas en relación a lo estudiado en los dos cursos anteriores, dado que se involucra al experto de datos en la resolución del problema. En la segunda parte se cubren conceptos de modelado y simulación de sistemas. Los modelos de simulación representan una herramienta de predicción efectiva cuando se conoce la mecanística del sistema a modelar pero se tiene incertidumbre sobre los datos de entrada. Finalmente, en la tercera parte se estudian métodos orientados a desarrollar sistemas de recomendación. En este caso, además de estudiar conceptos teóricos sobre las técnicas típicamente utilizadas por los sistemas de recomendación en diferentes dominios, se adquirirá conocimiento práctico aplicando algunas de estas técnicas en la implementación de un sistema básico de recomendación.

Objetivos

Con este curso se espera que el alumno logre:

- identificar las situaciones en las que la exploración visual y la minería de datos interactiva permiten superar a los métodos de minería de datos completamente automatizados;
- reconocer las metáforas visuales más apropiadas en función del tipo de dato;
- desarrollar conocimientos prácticos para implementar visualizaciones de datos interactivas;
- aplicar métodos de bondad de ajuste de diferentes funciones probabilísticas a datos de entrada;
- reconocer qué tipo de simulación corresponde aplicar dependiendo del sistema;
- conocimientos prácticos para realizar y evaluar simulaciones estáticas y dinámicas;
- familiarizarse con distintas técnicas aplicadas en el área de sistemas de recomendación, tales como el modelado del usuario y la factorización de matrices;
- familiarizarse con algoritmos de recomendación basados en contenido y de filtrado colaborativo;
- adquirir conocimientos básicos sobre otros tipos de sistemas de recomendación, tales como híbridos, demográficos y basados en conocimiento;
- aplicar los conceptos adquiridos en el desarrollo de un sistema básico de recomendación utilizando herramientas de software disponibles;
- conocer las métricas y técnicas que permiten evaluar a los sistemas de recomendación.

Contenidos Mínimos

- **Visualización de datos.** Visualización de datos versus visualización científica. Buenas prácticas en la visualización de datos. Metáforas visuales. Visualización interactiva. Librerías de visualización.
- **Minería de datos interactiva.** Incorporación del usuario experto en el ciclo de aprendizaje (analítica visual). Guia para el desarrollo de herramientas de analítica visual. Analítica de a pares. Revisión de aplicaciones de analítica visual.
- **Análisis de la entrada.** Ajuste de datos a funciones de distribución de probabilidad. Técnicas generativas de datos.
- **Simulación.** Simulación de Monte Carlo. Análisis de sensibilidad. Vistas del mundo. Simulación de eventos discretos. Verificación y validación de modelos.
- **Sistemas de recomendación:** Modelado de usuarios. Factorización de matrices. Recomendación basada en contenidos. Filtrado colaborativo. Sistemas de recomendación

híbridos, demográficos y basados en conocimiento. Confiabilidad en sistemas de recomendación. Herramientas (e.g. LensKit o similares). Evaluación.

Actividades

La práctica se organiza en tres partes. La primera corresponde a los temas de analítica visual, donde se plantearán distintas problemáticas y estrategias visuales para resolver tales problemáticas. Los alumnos discutirán y evaluarán en forma crítica, ventajas y desventajas de las distintas propuestas. Algunos ejemplos concretos de distintos escenarios son: búsqueda exploratoria en datos multivariados, visualización análitica de textos y ajustes de modelos de aprendizaje no supervisado. Asimismo, se brindará un tutorial para desarrollar habilidades prácticas sobre una o más librerías de visualización (e.g., Gephi, Bokeh, d3, etc).

La segunda corresponde al modelado de variables aleatorias y simulación. Para el modelado de variables aleatorias se presentarán varios conjuntos de muestras de datos correspondientes a distintas variables a modelar mediante la generación aleatoria de datos, y los alumnos deberán realizar visualizaciones y tests de bondad de ajuste para identificar las distribuciones de dichas variables. Para estos ejercicios se utilizará como tecnología Google Analytics, que es un complemento de Google Sheets. En lo referido a simulación, se presentarán varios casos de estudio que los alumnos deberán modelar mediante técnicas de simulación estática y dinámica. En la generación de los modelos, se hará énfasis en cómo diseñar los experimentos en términos de los objetivos de la simulación, y en cómo analizar la salida usando distintas estrategias de visualización de los resultados (uso de gráficos de sensibilidad y de frecuencia) e interpretando las métricas estadísticas reportadas por la simulación. Para estos ejercicios se utilizará como tecnología Risk Solver, que es un complemento de Google Sheets.

La tercera parte de las actividades prácticas corresponde al estudio de sistemas de recomendación. En base a los conceptos fundamentales cubiertos en las clases teóricas, se realizarán ejercicios básicos diseñados para fijar conceptos abstractos e ilustrarlos con ejemplos aplicados. Finalmente, se construirán y evaluarán sistemas de recomendación utilizando distintas herramientas (tales como el paquete Surprise de Python) y conjuntos de datos públicos como MovieLens.

Para todas estas actividades prácticas, la presentación inicial de las consignas y lineamientos generales que guían cada ejercitación, como también la posterior discusión y puesta en común de los resultados alcanzados, se realizará durante las clases prácticas presenciales. Mientras que la resolución en sí de las actividades se efectuará a distancia con los docentes supervisando a los alumnos mediante distintas herramientas tecnológicas (foros y chats en Moodle).

Bibliografía

- Aggarwal, C. C. (2018). Recommender Systems: the textbook. Springer International Publishing.
- Aggarwal, C. C. (2016). An Introduction to Recommender Systems. In Recommender Systems (pp. 1–28). Cham: Springer International Publishing.
- Aigner, W., Miksch, S., Schumann, H., & Tominski, C. (2011). Visualization of Time-Oriented Data. London: Springer London.

- Banks, J. (Ed.). (1998). Handbook of simulation: principles, methodology, advances, applications, and practice. John Wiley & Sons.
- Brath, R., & Jonker, D. (2015). Graph analysis and visualization: discovering business opportunity in linked data. Hoboken: John Wiley & Sons Inc.
- Evans, J. R., & Olson, D. L. (2001). Introduction to simulation and risk analysis. Prentice Hall PTR.
- Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1), 5–53.
- Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). Recommender Systems: An Introduction. Cambridge University Press.
- Keim, D., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). Mastering the Information Age Solving Problems with Visual Analytics. Eurographics Association.
- Munzner, T., & Maguire, E. . (2014). Visualization Analysis & Design. CRC Press.
- Resnick, P., & Varian, H. R. (1997). Recommender systems. Communications of the ACM, 40(3), 56–58.
- Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. Computer, 42(8), 30–37.
- Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender Systems: Introduction and Challenges. In Recommender Systems Handbook (pp. 1–34). Boston, MA: Springer US.
- Sun, G.-D., Wu, Y.-C., Liang, R.-H., & Liu, S.-X. (2013). A Survey of Visual Analytics Techniques and Applications: State-of-the-Art Research and Future Challenges. Journal of Computer Science and Technology, 28(5), 852–867.
- Wills, G. (2012). Visualizing time: designing graphical representations for statistical data. Springer Verlag.
- Wu, Y., Cao, N., Gotz, D., Tan, Y.-P., & Keim, D. A. (2016). A Survey on Visual Analytics of Social Media Data. IEEE Transactions on Multimedia, 18(11), 2135–2148.

Cabe destacar que toda la bibliografía indispensable para la realización del curso estará disponible online o en la Biblioteca del DCIC.

5.7 Big data y computación de alto rendimiento

El análisis de datos sobre grandes cantidades de datos—*big data*—da lugar a nuevos desafíos en el sentido que técnicas tradicionales dejan de ser tratables computacionalmente. Esta situación es agravada en las situaciones en donde se requieren respuestas en corto tiempo. Otro escenario complejo es en situaciones en donde los datos no son estáticos sino un flujo de datos continuo y potencialmente infinito (*data streams*) y por ende no es posible operar con todos los datos en memoria al mismo tiempo.

Para atacar dichos inconvenientes, se recurre a dos estrategias. La primera es una adaptación que apunta a reducir la complejidad algorítmica y la cantidad de datos que se analizan, aun cuando esto signifique un detrimento en la precisión de los resultados. La segunda estrategia corresponde al uso de técnicas de programación que sacan provecho del cómputo en paralelo. Para que el científico de datos pueda utilizar estas últimas técnicas, es importante que las mismas provean mecanismos automáticos de optimización del balance de carga, optimización de la transferencia de datos y recuperación ante fallas.

Objetivos

Con este curso se espera que el alumno logre:

- identificar casos en donde los algoritmos clásicos fallen en dar soluciones satisfactorias;
- conocer variantes sobre los algoritmos clásicos que mejoran la complejidad de tiempo del peor caso;
- analizar situaciones de compromiso que contemplen la precisión deseada y el tiempo de ejecución necesario entre dos o más propuestas de solución;
- reconocer las técnicas principales de distribución de cómputos;
- aplicar diseños de soluciones basados en el modelo MapReduce;
- comprender los requerimientos necesarios para procesar datos de tipo stream;
- desarrollar conocimientos prácticos para analizar datos de tipo stream.

Contenidos Mínimos

- **Análisis y aprendizaje automático en grandes datos**. Problemas y desafíos en big data. Situaciones de compromiso: precisión vs. tiempo de ejecución. Variantes eficientes para algoritmos clásicos (e.g. vecinos más cercanos, máquinas de soporte vectorial).
- **Computación distribuida**. Arquitecturas de clusters. Sistemas de archivos distribuidos (GFS y HDFS). MapReduce y Hadoop. Utilización de modelos basados en MapReduce desde bases de datos (e.g. MongoDB) y otros lenguajes de programación (e.g. Python).
- **Datos infinitos y no estacionarios**. Modelo de datos de tipo *stream*. Aprendizaje por lotes vs en línea (*batch vs online*). Sampleo, filtrado y análisis de streams. Spark y Spark streaming.

Actividades

El programa de este curso apunta al planteo, discusión y puesta en práctica de situaciones que por su gran volumen o velocidad de cambio resultan problemáticas y que, a pesar de darse con frecuencia en situaciones reales, no se describen en los libros de texto. Las situaciones problemáticas pueden surgir debido a que el volumen de datos a analizar es demasiado grande para ser cargados en memoria o porque el problema se vuelve computacionalmente intratable. Las clases teóricas, a ser dictadas en un laboratorio de computación, plantearán situaciones en la que se muestra cómo los grandes datos imposibilitan el tratamiento de un problema utilizando herramientas clásicas. En este sentido, se discutirán dos enfoques diferentes. El primero es el de la mejora de los algoritmos en cuanto a su orden de ejecución a expensas de la pérdida de precisión. El segundo enfoque, y sobre el cual se pondrá más énfasis, apuntará a presentar modelos de cómputo distribuido, los cuales son comúnmente usados dentro de la ciencia de datos. Por otra parte, las clases prácticas permitirán experimentar utilizando modelos de cómputo distribuido. Para ello se trabajará sobre un cluster Hadoop en donde se podrá experimentar con el modelo de programación MapReduce. Además, se propondrán datos reales de stream (e.g. API de Twitter) para que sean analizados con herramientas específicas para tal fin (e.g. Spark streaming).

Las actividades prácticas se realizarán tanto en forma presencial en un laboratorio de computación como a distancia mediante acceso remoto usando Google Colab u otras herramientas. La presentación inicial de las actividades y problemas a realizar a distancia, como la posterior discusión y puesta en común de los resultados, se realizará durante las clases prácticas presenciales.

Además, los docentes supervisarán a los alumnos mediante distintas herramientas tecnológicas (emails, foros, etc).

Para abordar estos dos enfoques para el tratamiento de grandes volúmenes de datos se propone desarrollar las siguientes actividades prácticas:

A) Experimentar con el uso de herramientas de acceso remoto y colaborativas (por ejemplo, programación en Python usando Google Colab). Actividad presencial

B) Resolver ejercicios que permitan comprobar la relación entre el orden de ejecución y la precisión de los algoritmos (por ejemplo, la simulación Barnes-Hut para el problema de los n cuerpos o locality-sensitive hashing para la búsqueda del vecino más cercano). Actividad no presencial.

C) Aprender a acceder y usar un cluster para cómputo distribuido (por ejemplo, el modelo de programación MapReduce sobre un cluster Hadoop). Actividad presencial.

D) El uso de APIs para acceder a datos reales en streaming y herramientas para su análisis (por ejemplo, el uso de Spark streaming para procesar datos de Twitter). Actividad presencial.

E) Implementar la solución a un problema simple que requiera integrar las herramientas de software y técnicas presentadas anteriormente (por ejemplo, aplicar TF-IDF en tiempo real sobre una ventana en el tiempo de tweets). Actividad no presencial.

F) Puesta en común y discusión de la solución al problema planteado y los resultados obtenidos. Actividad presencial.

Bibliografía

- Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107.
- Ghemawat, S., Gobioff, H., Leung, S.-T., Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. In Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles - SOSP '03 (Vol. 37, p. 29). New York, New York, USA: ACM Press.
- Joachims, T., & Thorsten. (2006). Training linear SVMs in linear time. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06 (p. 217). New York, New York, USA: ACM Press.
- Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of Massive Datasets (2nd ed.). Cambridge University Press.
- Zaharia, M., Das, T., Li, H., Hunter, T., Shenker, S., & Stoica, I. (2013). Discretized streams. In Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles - SOSP '13 (pp. 423–438). New York, New York, USA: ACM Press.

Cabe destacar que toda la bibliografía indispensable para la realización del curso estará disponible online o en la Biblioteca del DCIC.

5.8 Gestión y presentación de proyectos de ciencia de datos

El desarrollo de un producto tradicional de software requiere de la planificación y ejecución cuidadosa de distintas etapas que van desde la elicitación de requerimientos a la entrega final,

pasando por otras etapas que comprenden el diseño, la implementación y su evaluación. Dichas etapas rara vez suceden en forma lineal, sino que se trata de un proceso iterativo en el que las etapas se repiten para dar lugar a cambios, mejoras y correcciones.

Un proyecto de ciencia de datos guarda cierta similitud con un proyecto de software tradicional, pero con la incertidumbre como característica central para el caso de un producto de ciencia de datos. De no existir incertidumbre, el científico de datos no tendría que explorar, hipotetizar, aplicar y evaluar distintos enfoques para resolver problemas. De igual manera, el científico de datos para realizar su trabajo se apoya fuertemente en la estadística, cuyo fundamento es la incertidumbre.

En este curso se trabajará sobre el ciclo de vida de un proyecto de ciencia de datos y sobre distintos aspectos importantes relacionados a la ejecución del mismo dentro de una organización, la cual puede no estar familiarizada con sus prácticas, ventajas y limitaciones. Finalmente, se discutirán aspectos éticos sobre el análisis de los datos, principios de igualdad y discriminación en el análisis de datos y la preservación de la anonimidad.

Objetivos

Con este curso se espera que el alumno logre:

- conocer y administrar las etapas de un proyecto de ciencia de datos y diferentes metodologías que soportan los ciclos de vida de proyectos;
- entender el rol que la ciencia de datos pueden desempeñar dentro de una organización;
- valorar la importancia de involucrar a distintos participantes dentro de una organización y buscar en forma continua preguntas que sustenten el análisis de los datos;
- reconocer cómo presentar proyectos de ciencia de datos en función de la audiencia destinada;
- identificar aspectos éticos en el uso y procesamiento de datos;
- conocer técnicas básicas para la anonimización de datos.

Contenidos Mínimos

- **Ciclo de vida de un proyecto de ciencia de datos.** Preparación, diseño, implementación, evaluación y entrega. Adaptación del producto a la audiencia. Ingeniería inversa del proceso analítico.
- **Ciencias de datos dentro de una organización.** Involucración amplia de los participantes. Valorización de los datos internos como un activo de la organización. Cambio de paradigma para la toma de decisiones. Roles del profesional de ciencia de datos.
- **Ética en ciencia de datos.** Principio de “mínimo necesario”. Sesgo y discriminación en el análisis de datos. Anonimización de datos: k-anonimizaciones.

Actividades

En este curso el alumno deberá abstraerse de los aspectos técnicos de un proyecto de ciencia de datos y analizarlo desde el entorno donde será llevado a cabo, por lo que aprenderá a administrar las etapas del ciclo de vida en el contexto de una organización. La teoría, dictada en un laboratorio de computación, expondrá las características salientes del ciclo de vida de un proyecto

de software. Se enseñarán las metodologías CRISP-DM, Modelo de Cascada, y SCRUM, considerando distintos enfoques para el desarrollo – lineal, incremental, iterativo, y adaptativo. A fin de incorporar conocimientos sobre las metodologías, se presentarán diferentes casos de estudio. Los alumnos trabajarán en equipos. Cada equipo abordará uno de los casos, que le será asignado por el docente, y evaluará las ventajas y desventajas de adoptar cada una de las metodologías y seleccionará la más adecuada. Las conclusiones serán presentadas por cada equipo y discutidas entre todos los alumnos, haciendo una síntesis común en una actividad taller a desarrollarse en el aula.

En base al caso asignado, los equipos estudiarán las tareas y posibles instrumentos a usar en cada una de las etapas básicas de un proyecto – preparación, diseño, implementación, evaluación y entrega. Para la preparación, deberán adquirir capacidades para comprender los objetivos del proyecto y hacer una exploración identificando y evaluando posibles fuentes de datos, así como también identificar técnicas para la minería y limpieza de los mismos. Para el diseño, deberán presentar el modelo de datos propuesto y los algoritmos a aplicar para realizar el análisis. Para la implementación, se espera que planifiquen el desarrollo de un prototipo simple. Para la evaluación, deberán sintetizar cómo la solución planteada satisface los objetivos del negocio y las posibles extensiones, generalizaciones y limitaciones. Finalmente, cada equipo definirá el contexto y condiciones necesarias para la entrega de la solución.

Se espera que este trabajo sea desarrollado por los alumnos de manera gradual durante el cursado de la materia. Para ello, se espera que haya discusiones en foros (por equipo y para todo el curso) habilitados en la plataforma en línea. Al finalizar el desarrollo, los alumnos deberán presentar de manera presencial y en forma oral el proyecto desarrollado para todos los alumnos y docentes. Asimismo, deberán presentar en forma digital un documento definiendo la planificación del proyecto, y la justificación de las decisiones tomadas. En estas actividades, se pondrá especial énfasis en la etapa de elicitación de requerimientos, las cuales son de naturaleza exploratoria dentro de la ciencia de datos, y la presentación de los resultados los cuales deben adaptarse a la audiencia destinada y en la justificación de los resultados alcanzados.

Durante el desarrollo de la materia, se realizarán discusiones grupales sobre la implementación de la ciencia de datos dentro de una organización, así como también sobre aspectos éticos en el almacenamiento y procesamiento de datos personales. Como parte de los contenidos prácticos se evaluarán distintos casos en donde se plantea su implementación hipotética como proyecto de ciencia de datos. También se trabajará con distintas técnicas prácticas de anonimización de datos. El taller de proyecto final integrador servirá como otro valioso mecanismo para la puesta en práctica de los conceptos analizados en este curso.

Las actividades prácticas se realizarán tanto en forma presencial en un laboratorio de computación como a distancia mediante acceso remoto usando Google Drive u otras herramientas. La presentación inicial de las actividades y problemas a realizar a distancia, como la posterior discusión y puesta en común de los resultados, se realizará durante las clases prácticas presenciales. Además, los docentes supervisarán a los alumnos mediante distintas herramientas tecnológicas (emails, foros, etc).

Bibliografía

- Aggarwal, C. C., & Yu, P. S. (2008). A General Survey of Privacy-Preserving Data Mining

Models and Algorithms (pp. 11–52). Springer, Boston, MA.
- Azevedo, A. and Santos, M.F. (2008) KDD, SEMMA and CRISP-DM: A Parallel Overview. Proceedings of the IADIS European Conference Data Mining, Amsterdam, 24-26 July 2008, 182-185.
- Bayardo, R. J., & Agrawal, R. (2005). Data Privacy through Optimal k-Anonymization. In 21st International Conference on Data Engineering (ICDE'05) (pp. 217–228). IEEE.
- Godsey, B. (2017). Think like a data scientist: tackle the data science process step-by-step. O'Reilly.
- Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic Bias. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 (pp. 2125–2126). New York, New York, USA: ACM Press.
- Hajian, S., & Domingo-Ferrer, J. (2013). A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. IEEE Transactions on Knowledge and Data Engineering, 25(7), 1445–1459.
- Marbán, O., Segovia, J., Menasalvas, E. & Fernández-Baizán, C. (2009). Toward data mining engineering: A software engineering approach. Information Systems, 34(1), 87–107.
- Moro, S., Laureano, R. & Cortez, P. (2011). Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.
- Provost, F., & Fawcett, T. (2013). Data science for business. O'Reilly.
- Schermer, B. W. (2011). The limits of privacy in automated profiling and data mining. Computer Law & Security Review, 27(1), 45–52.
- Wysocki, R.K. (2013) Effective Project Management: Traditional, Agile, Extreme. Wiley.
- Xindong Wu, Xingquan Zhu, Gong-Qing Wu, & Wei Ding. (2014). Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1), 97–107.

Cabe destacar que toda la bibliografía indispensable para la realización del curso estará disponible online o en la Biblioteca del DCIC.

5.9 Campos de aplicación y taller de proyecto final

Los distintos campos de aplicación a los que la ciencia de datos pueden aportar soluciones cuentan con características distintivas, tanto en términos de formato de los datos comúnmente utilizados, problemas típicamente planteados, tecnologías aplicables y soluciones esperadas. Este curso ofrecerá al alumno un panorama general sobre diferentes campos de aplicación de la ciencia de datos, tales como Bioinformática, Gobierno Abierto, Econometría, Minería de la Web y Ciberseguridad.

Una vez conocidos los campos de aplicación, el alumno tendrá la oportunidad de identificar una problemática real dentro de alguno de ellos que requiera una solución de ciencia de datos. Para el desarrollo de la solución, el alumno realizará un proyecto que integrará los conocimientos adquiridos durante la especialización. Esto permitirá que el alumno se ejercite en la aplicación de las tecnologías y metodologías estudiadas dentro de un dominio específico.

Objetivos

Con este curso se espera que el alumno logre:

- conocer los campos de aplicación y casos prácticos en los que las tecnologías actuales de ciencia de datos pueden ser utilizadas;
- conocer algunos de los problemas típicamente planteados en distintos dominios de aplicación;
- identificar una problemática real que requiera una solución de ciencia de datos;
- proponer una solución a la problemática identificada integrando conocimiento adquirido durante la carrera
- adquirir experiencia en la aplicación concreta de los conceptos y tecnologías estudiadas dentro de un dominio específico.

Contenidos Mínimos

- **Bioinformática.** Datos biológicos. Bases de datos: formatos y anotaciones. Herramientas de recuperación de datos (e.g., Entrez, DBGET y SRS). Herramientas para alineamiento de secuencias (e.g., FASTA y BLAST).
- **Gobierno abierto.** Datos abiertos: conceptos y teoría. Visualización y herramientas de análisis para datos abiertos. Infraestructuras de datos abiertos. Metadatos y ontologías para datos abiertos. Sensibilidad de datos. Datos abiertos para la formulación de políticas.
- **Econometría.** Modelos estructurales y especificación. Métodos de predicción clásicos en Econometría. Predicción en series de tiempo (in-sample, out-sample). Modelos de causalidad. Causalidad de Granger. Redes Bayesianas. Predicción y toma de decisiones.
- **Minería de la Web.** Técnicas de crawling, indexación y búsqueda en la Web. Análisis de contenido. Análisis de estructura. Análisis de uso. La Web invisible. La Web oscura. La Web semántica. Análisis de redes sociales y comunidades en la Web.
- **Ciberseguridad.** Clasificación de malware. Detección de botnets, actividad anómala en redes, transacciones fraudulentas y otras amenazas. Aprendizaje automático adversarial: Ataques por evasión y por envenenamiento.

Actividades

En la primera parte del curso, un grupo de docentes con experiencia en una variedad de campos de aplicación brindará una introducción general a un conjunto de problemas que se pueden abordar desde la ciencia de datos. Esta primera actividad se desarrollará analizando ejemplos específicos de problemas y diferentes soluciones propuestas, con un foco particular en detallar cómo fueron abordados los diferentes aspectos cubiertos en los cursos anteriores.

En una segunda etapa, cada alumno deberá elegir primero un campo de aplicación y luego un problema específico para la realización de un proyecto de ciencia de datos. Dependiendo de la temática elegida por cada alumno, los docentes a cargo podrán invitar a otros docentes de la Especialización a guiar el desarrollo de proyectos que requieran el dominio de un campo en particular. Una vez elegido el campo y problema, se llevarán a cabo encuentros presenciales entre cada alumno y el docente a cargo de la supervisión para resolver las consultas y realizar un seguimiento de las actividades. El desarrollo del proyecto contemplará pre-entregas en forma remota a fin de asegurar el avance del alumno y poder atender a tiempo los problemas que surjan.

Todas las actividades prácticas se realizarán tanto en forma presencial (en un laboratorio de computación) como a distancia; para la modalidad a distancia, se emplearán las herramientas comúnmente usadas en la disciplina. Para el acceso colaborativo a código fuente, ejemplos actuales de herramientas incluyen Google Colab y GitHub. Por otro lado, para la comunicación no presencial, los docentes podrán utilizar las herramientas que más se adapten a las necesidades de cada alumno y su proyecto, tales como correo electrónico, foros, chat, videoconferencia, etc.

Por último, la evaluación final del curso se realizará a través de la presentación de un informe documentando todas las actividades realizadas, resultados obtenidos y dificultades que se encontraron durante el desarrollo del proyecto.

Bibliografía

- Chakrabarti, S. (2003). Mining the Web: discovering knowledge from hypertext data. Morgan Kaufmann.
- Croft, B., Metzler, D., & Strohman, T. (2010). Search Engines Information Retrieval in Practice. Addison Wesley.
- Drăghici, S. (2008). Bioinformatics databases: design, implementation, and usage. CRC Press.
- Goodfellow, I., McDaniel, P., & Papernot, N. (2018). Making machine learning robust against adversarial inputs. Communications of the ACM, 61(7), 56–66.
- Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information systems management, 29(4), 258-268.
- Lesk, A. (2014). Introduction to bioinformatics. Oxford University Press.
- Letovsky, S. I. (Ed.). (2006). Bioinformatics: databases and systems. Springer Science & Business Media.
- Tauberer, J. (2012). Open government data.
- Ubaldi, B. (2013). Open government data: Towards Empirical Analysis of Open Government Data Initiatives.
- Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3-28.
- Vorobeychik, Y., & Kantarcioglu, M. (2018). Adversarial Machine Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 12(3), 1–169.
- Wooldridge, J. M. (2015). Introductory econometrics: A modern approach. Nelson Education.

Cabe destacar que toda la bibliografía indispensable para la realización del curso estará disponible online o en la Biblioteca del DCIC.