Ciencia de Datos: Un Enfoque Práctico en la Era del Big Data

Contenidos de Teoría (13h)

1. Ciencia de Datos, analítica avanzada y *big data* (1h)
    - La era de los datos, científico de datos, minería de datos y KDD, aplicaciones
    - Interpretabilidad vs. Precisión
    - Big Data: las tres 'V'. Presencia en medios e impacto económico
    - Casos reales: sistemas de recomendación en Amazon y Netflix, transacciones de tarjetas de crédito, epidemiología, redes sociales
    - Software: Weka, KNIME, KEEL, scikit-learn (Python), R…
    - Competición Kaggle
    - Hacia dónde vamos, errores comunes, tecnología emergente, evolución de popularidad
2. Análisis exploratorio de datos: visualización (1h)
    - Motivación, ejemplos de visualización de datos
    - Tipos de datos
    - Visualizaciones en datos categóricos, temporales, espaciales, multivariable, distribuciones…
    - Software: Tableau, QlikView, R, JavaScript…
3. Fundamentos de clasificación: árboles de decisión, *lazy*, RNA, bayesianos, evaluación (2h)
    - Definición del problema
    - Etapas en el proceso de clasificación
    - Clasificadores basados en instancias: lazy learning y vecinos cercanos
    - Árboles de decisión e inducción de reglas: ID3 y C4.5
    - Redes neuronales: MLP, RBFN
    - Métodos bayesianos: Naïve Bayes,
    - Evaluación: métricas, métodos, comparación, ROC
4. Preprocesamiento: selección y procesado de instancias y características, tratamiento del ruido (2h)
    - Importancia del preprocesamiento
    - Integración, limpieza, normalización, transformación…
    - Valores perdidos
    - Datos con ruido
    - Selección de variable e instancias
    - Discretización: CAIM
5. Clasificación avanzada: SVM, *ensemble learning*, problemas no balanceados, *deep learning* (2,5h)
    - Máquinas de soporte vectorial: SMO, LibSVM, kernels
    - Ensemble learning: bagging, boosting y binarización, AdaBoost, OVO, OVA
    - Clasificación no balanceada: introducción, soluciones, problemas, datasets, SMOTE
    - Deep learning, auto-encoder, Google DeepMind
6. Segmentación y relaciones: *clustering* y reglas de asociación (2h)
    - Aprendizaje no supervisado
    - Clustering: definición, análisis, medidas de distancia y similaridad, métodos de particionamiento, k-means, métodos jerárquicos aglomerativos y divisivos

- Reglas de asociación: market basket analysis, conceptos, asociaciones booleanas y cuantitativas, unidimensionales y multidimensionales, soporte y confianza, algoritmo Apriori, medidas de interés, lift, otros algoritmos (FP-Growth, OPUS, QAR).

7. Aprendizaje incremental y *data stream mining* (1h)
    - Problemas incrementales, aprendizaje dinámico, incremental vs. flujo de datos
    - Aprendizaje incremental: concepto, criterios de evaluación, modos de aprendizaje incremental de ejemplos, clases y atributos, aplicabilidad, construcción incremental de árboles de decisión, árboles Hoeffding, VFDT, clustering incremental, STREAM
    - Minería de flujo de datos: definición, tareas de clasificación, agrupamiento y patrones frecuentes, concept drift, clasificación, CVFDT, ventana deslizante y factor de decaimiento, ADWIN, ensemble learning para data stream, detección de concept drift, clustering con cambio de concepto, CluStream, patrones frecuentes, heavy hitters, association stream mining, ejemplos reales, líneas abiertas
    - Software: RapidMiner, MOA, Sofia-ML, Flink, streamDM…
8. *Big data*: fundamentos y paradigmas (1,5h)
    - Las 3+5 'V'
    - Paradigma MapReduce
    - Ecosistema Hadoop, Spark
    - Mahout, MLlib, FlinkML, H_2O…
    - Big data preprocessing
    - Bibliotecas para analítica de datos, casos de estudio
    - Aplicaciones reales

Contenidos de Prácticas (17h)

1. KNIME (5,5h): predicción fundamental
    - Introducción
    - Instalación de KNIME
    - Introduccion a la interfaz gráfica de KNIME: editor, nodos, manipulación de datos, visualización, hiliting, etc.
    - Creación de un flujo de dato básico (WorkFlow)
    - Ejercicios básicos
    - Scripting in KNIME
    - Validación cruzada
    - Preprocesamiento en KNIME: selección de características, normalización, selección de prototipos, discretización, valores perdidos (missing values)
    - Técnicas de comparación entre métodos
    - Ejercicios avanzados
2. Python para Ciencia de Datos (6,5h): visualización y predicción avanzada
    - Introducción a Python y la interfaz de Jupyter. Tipos de datos, carga de datos de distintas fuentes.
    - Llamando a R y C++ desde Python.
    - Análisis exploratorio de datos con Pandas.
    - Exploración visual de los datos.
    - Introducción del paquete Scikit-Learn para aprendizaje.
    - Diseño de experimentos de aprendizaje supervisado.
    - Particionamiento de datos, entrenamiento y análisis de modelos de aprendizaje.

- Ejercicios prácticos con SVMs, Boosting, Bagging, Random Forest y otros.
- Visualización y comparación de resultados para selección de modelos.
- Minería de reglas de asociación.
- Ejercicios prácticos de segmentación con reglas de asociación.
- Introducción al agrupamiento de datos.
- Visualización de los clusters y sus características fundamentales.
- Ejercicios de clustering básico, jerárquico, por densidad y basado en modelos.

3. Spark + MLLib (5h): big data
    - Breve introducción de la parte práctica de Spark y MLLib.
    - Instalación de Spark de manera local.
    - Ejecución de ejemplos prácticos para SparkR.
    - Introducción a la plataforma cloud para Spark: DataBricks Community (DB).
    - Ejecución de operaciones básicas en DB (transformaciones y acciones).
    - Ejercicio guiado para implementar WordCount usando DB.
    - Ejercicio guiado para entender MLLib y su funcionalidad usando DB.

Trabajo Individual (45h)

- Estudio de la materia y elaboración de respuestas a varias preguntas teóricas formuladas para cubrir los distintos temas estudiados.
- Participación en una competición de Kaggle (https://www.kaggle.com/) o DrivenData (https://www.drivendata.org/). Durante varias semanas el alumno pone en práctica todos los conocimientos adquiridos en el curso con la ayuda de las herramientas software estudiadas para abordar un problema real de ciencia de datos en un entorno competitivo mundial que les anima a dar lo mejor de sí.