Especialización en Analítica y Ciencia de Datos
Universidad de Antioquia. Facultad de Ingeniería

Contenido resumido de las materias

Semestre I

---

01. Fundamentos de programación para ciencia de datos.

1. Jupyter Notebook System sobre SPARK
2. Tipos de datos , lectura de archivos en diversos formatos y Funciones en Python
3. Librerías numéricas de python.
4. Dataframe. Creación, uso y transformación.
5. El SQL y la manipulación de dataframe.
6. Generación de números aleatorios, muestras y distribuciones.
7. Visualización de datos con matplotlib

02. Estadística y análisis exploratorio.

Unidad 1: Introducción variables aleatorias y momentos

- Variables aleatorias, función de distribución, distribuciones bivariadas, marginales, condicionales. Teorema de Bayes.
- Esperanza de una variable aleatoria, momentos, media y mediana, covarianza y correlación, esperanza condicional. Distribuciones especiales.
- Tipos de variables y muestreo. Muestreo de funciones de distribución. Taller sobre muestreo, gráficas: histograma y estimador de densidad kernel.

Unidad 2: Función de distribución Gausiana y teorema de Bayes

- Distribuciones a priori y a posteriori, distribuciones conjugadas, estimador de Bayes.
- Función de Distribución Gaussiana Multivariada, distribución Gausiana Condicional, distribución Gausiana Marginal, Inferencia, Máxima verosimilitud para la Gaussiana.

Unidad 3: Intervalos de confianza y test de diferenciación de medias

- Distribución chi-cuadrada, distribución *t*, Intervalos de confianza,
- Pruebas de hipótesis, el test *t*, test de bondad de ajuste, KS-test, ANOVA y MANOVA.
- Taller práctico sobre estimación de intervalos de confianza y test de hipótesis. Gráfica volcano.

Unidad 4: Preparación de datos

- Preparación de datos, atributos redundantes, limpieza de datos y normalización.
- Detección de datos atípicos, imputación de variables y codificación.
- Taller de aplicación preparación de datos y visualización.

03. Programación sobre grandes volúmenes de datos.

1. Sistema de archivos distribuidos (HDFS)

2. Map-reduce
3. Eficiencia de los procesos paralelos y distribuidos.
4. SPARK y RDD
5. Dataframe y SQL.
6. Las librerías básicas para ML.

04. Aprendizaje automático I.

Unidad 1: Introducción y fundamentos del aprendizaje automático

- Introducción, Definiciones, Sklearn Script básico de una simulación en ML
- Regresión lineal y regresión logística + Taller
- Taller con dataset grande limpieza de datos + train/test con métrica de score básica para regresión y para clasificación

Unidad 2: Clasificación y selección de modelos

- Paramétrico vs No paramétrico: K-nn vs Gaussian. Taller sobre los modelos, fronteras de decisión.
- Selección de modelos, overfitting y regularización.
- Taller con dataset real selección de modelos: k-fold, k-folds estratificado, k-fold por grupos, Boostrapping.

Unidad 3: Árboles de decisión y máquinas de vectores de soporte

- Árboles, Bagging + Random Forest.
- Máquinas de Vectores de Soporte, One vs All, All vs All
- Taller práctico comparación de modelos de la semana

Unidad 4: Boosting y selección de características

- Boosting: Adaboost y Gradient Boosting
- Selección de características e importancia de variables
- Taller de aplicación de las técnicas de la semana

05.Seminario

1 LOS FUNDAMENTOS METODOLÓGICOS Y CONCEPTUALES PARA LA IDENTIFICACIÓN Y FORMULACIÓN DE PROYECTOS DE INVESTIGACIÓN APLICADA, A NIVEL DE ANALÍTICA Y CIENCIA DE DATOS.
2 CONSIDERACIONES PRÁCTICAS DE CONCEPTOS DE ANALÍTICA DE DATOS EN PROYECTOS APLICADOS
3 ELABORACIÓN Y PRESENTACIÓN DE PROPUESTA MONOGRÁFICA.

Semestre II

06. Aprendizaje Automático II

Unidad 1: Fundamentos de clustering y reducción de dimensionalidad
- Clustering k-means, agglomerative clustering, silhouette
- PCA, LDA, t-SNE
- Taller uso de técnicas y visualización de cluster en baja dimensión

Unidad 2: Técnicas avanzadas e integración en flujos de trabajo
- Spectral clustering
- Pipelines

Unidad 3: Aprendizaje por refuerzo
- Reinforcement Learning

Unidad 4: Taller
- Taller de aplicación integral con datasets propios del estudiante.

07. Deep Learning

Unidad 1: Modelos derivados de los datos y diseño de algoritmos de machine learning
Unidad 2: Introducción a las redes neuronales artificiales y a las redes profundas
Unidad 3: Introducción a Tensorflow
Unidad 4: Analítica de imágenes con redes neuronales convolucionales
Unidad 5: Analítica de series temporales con redes neuronales recurrentes

08. Data streaming y servicios en la nube

Unidad 1. Introducción al Streaming de Datos (Clasificación native stream processing - Micro-batch processing , Tipos de Datos - JSON, XML, YAML, Protocol Buffers, Apache Thrift - Data lakes vs Data Stream)

Unidad 2. Tipos de Procesamiento (Procesamiento basado en eventos,Batch File-Based Processing, Continuous Operator Stream Processing, Stream Processing Services - Fuentes de Datos vs Almacenamiento de datos)

Unidad 3. Servicios y plataformas en la Nube

09. Visualización

1. Introducción
2. Conceptos sobre técnicas de visualización
3. Teoría de la Percepción
4. Teoría del Color
5. Teoría de análisis gráfico estadístico multivariado
6. Gramática de los Gráficos
7. Análisis Exploratorio de Datos
8. Taxonomía de Gráficos (*Tabla Periódica*).
    a. Distribución, Correlación,Barras, Jerárquicos, Redes, Evolución, etc.

9. Técnicas modernas de visualización de información. Enfoque hacia altos volúmenes de datos, con reducción de dimensión
10. Caso de estudio (*uso práctico de las técnicas y tecnologías presentadas en el curso*)
11. Análisis de ventajas y desventajas del uso de la Visualización

10. Aspectos éticos y jurídicos de la gestión de la información

1. Marco Constitucional de los Derechos Fundamentales
2. Régimen de la Protección de Datos Personales en Colombia y Europa
3. Privacidad por diseño, Seguridad de la información por defecto y Ética desde el diseño
4. Cumplimiento de la Seguridad de la Información
5. Guía de Responsabilidad Demostrada de la SIC
6. Gestión de riesgos tecnológicos y trascendencia jurídica
7. Responsabilidad derivada del desarrollo tecnológico
8. Gestión de las creaciones originadas en la ciencia de datos y tecnologías de la industria 4.0.
9. Equilibrio entre datos abiertos y datos privados