Elementos de Ciencia de Datos

I. Características del curso

1.1 **Curso bandera**

Este es el curso bandera del *Área de Concentración de Probabilidad e Inferencia Estadística en Ciencia de Datos* de la Maestría en Probabilidad y Estadística (MPyE). Se recomienda que sea un curso de tercer semestre de la MPyE.

1.2 **Prerrequisito**

El curso está diseñado para los alumnos que hayan llevado el programa de cursos de primer año de la MPyE, en especial los cursos de modelos estocásticos I y II, modelos estadísticos I e inferencia estadística I.

1.3 **Curso en paralelo deseable**

Es conveniente que se lleve en paralelo el curso de Modelos de probabilidad discretos a gran escala, en donde se vean temas de gráficas y redes.

II. Objetivos

Al final del curso el alumno**:**

2.1 Entenderá las principales problemáticas de los datos modernos y su análisis.

2.2 Dominará las técnicas estadísticas de agrupamiento y clasificación de datos masivos.

2.3 Conocerá elementos de aprendizaje máquina y minería de datos para el análisis de datos masivos.

2.4 Obtendrá las bases para desarrollar técnicas ralas/dispersas.

III. Contenido

3.1 **Panorama y naturaleza de datos masivos**: Fuentes de generación de datos masivos. Volumen, velocidad, variedad, veracidad y valor. Dificultades computacionales.

3.2 **Análisis visual de datos masivos:** Gráficas trellis, coordenadas paralelas, gráficas interactivas, tours, proyecciones.

3.3 **Métodos de clasificación**: Probabilidad de clasificación errónea y su costo (ECM), clasificación minimizando el ECM, análisis

discriminante, algoritmo de *k*-vecinos más cercanos, modelos de regresión, redes bayesianas, árboles de decisión.

3.4 **Métodos de búsqueda de estructuras y agrupamiento de datos:** Estimación no paramétrica de densidades**,** reducción de dimensión, análisis de componentes principales, factorización de matriz no negativa, análisis de factores, análisis de conglomerados, métodos espectrales, análisis fuzzy, modelos gráficos no-dirigidos I.

3.5 **Técnicas de aprendizaje máquina para el análisis de datos:** Métodos kernel, máquinas de soporte vectorial (svm's), algoritmo EM, modelos de mezclas gaussianas, redes neuronales.

3.6 **Matrices dispersas:** El Lasso, análisis de componentes principales para matrices dispersas, aproximaciones de rango uno, modelos gráficos no-dirigidos II.

3.7 **Temas selectos:** Datos masivos y diseño de experimentos, el problema *Big n vs Big p***,** inferencia causal.

IV. Bibliografía

4.1. Liu, S., McGree, J., Ge, Z. and Xie, Y. (2016) *Computational and Statistical Methods for Analysing Big Data with Applications*. Academic Press/Elsevier.

4.2. Koch, I. (2014) *Analysis of Multivariate and High-Dimensional Data*. Cambridge University Press.

4.3. Buhlmann, P., Drineas, P., Kane, M. and van der Laan, M. (2016) *Handbook of Big Data*. CRC/Chapman and Hall.

4.4. National Research Council (2013) *Frontiers in Massive Data Analysis*. The National Academies Press.

4.5. Ivezic, Z., Connolly, A.J., VanderPlas, J.T. and Gray, A. (2014) *Statistics, Data Mining, and Machine Learning in Astronomy*. Princeton University Press.

4.6. Izenman, A.J. (2008) *Modern Multivariate Statistical Techniques*. Springer.

4.7. Pourahmadi, M. (2013) *High-Dimensional Covariance Estimation*. Wiley.

4.8. Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer.