

## 3. Materias optativas maestría en PyE

### 3.1. INTRODUCCIÓN A CIENCIA DE DATOS

#### I. Objetivos

- Entender las principales problemáticas de los datos modernos y su análisis.
- Dominar las técnicas estadísticas de agrupamiento y clasificación de datos masivos.
- Entender y usar las técnicas asociadas a reducción de dimensionalidad y regularización.
- Conocer los elementos de aprendizaje máquina y minería de datos para el análisis de datos masivos.
- Obtener las bases para desarrollar técnicas raras/dispersas.

#### II. Contenido

##### 1. Métodos de clasificación

- 1.1. Clasificador óptimo de Bayes
- 1.2. Clasificador Bayesiano ingenuo
- 1.3. Análisis discriminante
- 1.4. Algoritmo de  $k$ -vecinos más cercanos
- 1.5. Modelos de regresión
- 1.6. Árboles de decisión
- 1.7. Modelos gráficos no-dirigidos I: Modelos gaussianos e independencia condicional

##### 2. Métodos de búsqueda de estructura

- 2.1. Descomposición en valores singulares (SVD)
- 2.2. Reducción de dimensión
- 2.3. Análisis de componentes principales (PCA)
- 2.4. Factorización no-negativa (NMF) y sistemas de recomendación
- 2.5. Elementos de álgebra lineal aleatorizada
- 2.6. Proyecciones aleatorias y el Lema de Johnson-Lindenstrauss

##### 3. Matrices dispersas

- 3.1. Regresión lasso
- 3.2. Modelos gráficos no-dirigidos II: Modelos gaussianos sparse

##### 4. Agrupamiento de datos

- 4.1. Análisis de conglomerados
- 4.2. Métodos espectrales para agrupamientos y caminatas aleatorias
- 4.3. Algoritmo EM
- 4.4. Modelos de mezclas gaussianas

##### 5. Técnicas de aprendizaje máquina para el análisis de datos

- 5.1. Métodos kernel
- 5.2. Máquinas de soporte vectorial
- 5.3. Redes neuronales

#### III. Prerequisitos

## Bibliografia recomendada

- [1] Afonso S. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. *Lecture Notes*, 2015.
- [2] C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 4 of *Information Science and Statistics*. Springer New York, 2016.
- [3] Peter Bühlmann, Petros Drineas, Michael Kane, and Mark van der Laan. *Handbook of Big Data*. CRC Press, 2016.
- [4] Peter Bühlmann and Sara Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.
- [5] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013.
- [6] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Elsevier, 2013.
- [7] Christophe Giraud. *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, 2021.
- [8] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009.
- [9] Michael W. Mahoney. Lecture notes on randomized linear algebra. *arXiv preprint arXiv:1608.04481*, 2016.
- [10] Deborah Nolan and Duncan Temple Lang. *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving*. CRC Press, 2015.
- [11] Mohsen Pourahmadi. *High-Dimensional Covariance Estimation*, volume 882. John Wiley & Sons, 2013.
- [12] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.