

CURSO

INTRODUCCIÓN A LA CIENCIA DE DATOS Y AL MACHINE LEARNING PARA CIENTÍFICOS SOCIALES

Instituto de Altos Estudios Sociales - UNSAM

Docente: Dr. Germán Rosati

PRESENTACION

Muchas áreas del conocimiento se han beneficiado de la aplicación creativa de las técnicas englobadas bajo el nombre de “Machine Learning” o “Aprendizaje Automático”. Sin embargo, las ciencias sociales argentinas no parecen haber hecho uso de las mismas. El curso se propone realizar una primera aproximación a dichas técnicas.

Se buscará un acercamiento que aborde un enfoque conceptual (fundamentos teórico-metodológicos, casos y problemas de aplicación, etc.) como técnico (análisis de algoritmos, herramientas con interfaces gráficas, etc.) de las mismas. Se realizarán lecturas y exposiciones teóricas sobre las temáticas, pero también se hará especial énfasis en su aplicación práctica, particularmente centrada en problemas relevantes de las Ciencias Sociales.

El curso puede darse en dos modalidades:

- Escribiendo código en R o Python, se hará énfasis en la necesidad de incorporar algunos fundamentos básicos de programación (estructuras de datos, loops, funciones, etc.) y se introducirán librerías para realizar procesamiento y modelado de información.
- Utilizando una herramienta de programación visual (“visual programming”) [<https://orange.biolab.si>] con el objetivo de poder centrar los esfuerzos en la dimensión conceptual y de aplicación de las herramientas y no en la parte de implementación/ programación/ coding de las mismas.

OBJETIVOS

El curso se propone que los asistentes:

- logren una primera aproximación a los problemas y discusiones epistemológicas y metodológicas acerca de la aplicación del machine learning en la actividad científica y las diferencias con el modelado estadístico tradicional
- incorporen fundamentos conceptuales del flujo de trabajo y de la metodología asociadas a estas técnicas
- conozcan y sepan implementar diversos algoritmos
- identifiquen situaciones de aplicación de este tipo de modelos a problemas de investigación básica y aplicada

MODALIDAD DE CURSADA

10 Clases de 3.5 hs.

Viernes de 14 a 17.30

PROGRAMA ANALITICO

Unidad 0. Fundamentos de programación en R o Python (2 clases) -opcional-

Estructura de datos: vectores, listas, dicts (Python), data frames. Estructuras de control: for, if. Uso e implementación de funciones ad-hoc. Importación y exportación de datos (.csv, .txt, .tab, .sav, etc.).

Librerías a trabajar en modalidad R: dplyr, ggplot, lubridate, caret, glmnet.

Librerías a trabajar en modalidad Python: pandas, seaborn, sklearn.

Unidad 1. ¿Qué es Machine Learning? (1 clases)

Aproximación general. Delimitación de los campos disciplinares en estudio: Data Science, Machine Learning e Inteligencia Artificial. Machine Learning, Ciencias y Ciencias Sociales. Problemas y desafíos epistemológicos y metodológicos en la actividad científica en general y particularmente haciendo énfasis en las Ciencias Sociales. Revisión de casos de aplicación

Unidad 2. Fundamentos conceptuales de Machine Learning (1 clase)

Flujo de trabajo: etapas, planteo del problema, extracción, limpieza, modelado y comunicación. Aprendizaje supervisado y no supervisado. Problemas de clasificación y regresión. Fuentes de error de un modelo, datos de entrenamiento y de testing, estimación del error de generalización (train test split, cross-validation) , balance sesgo-varianza. El “flujo de trabajo” en un proyecto de Machine Learning.

Unidad 3. Aprendizaje Supervisado (4 clases)

Modelos “simples” (no ensambles): Regresión lineal y logística. Evitando el overfitting: métodos de regularización (Ridge Regression y LASSO). Árboles de Decisión. Ensamblados de modelos: Bagging (Random Forest). Boosting (AdaBoost). Presentación de la herramienta Orange.

Unidad 4. Aprendizaje No Supervisado (1 clase)

Reducción de Dimensionalidad: Análisis de Componentes Principales. Clustering: K-Means, Clustering Jerárquico. Métricas de validación de Clustering.

Unidad 5. Taller práctico (1 clases)

Se abordará la resolución de un problema real recorriendo todo el proceso (problema-datos-limpieza-modelado-evaluación). Se introducirán algunas nociones

generales de Text Mining y Natural Language Processing, buscando 1) detectar tópicos relevantes en corpus de texto libre y 2) entrenar un modelo para clasificar dichos textos.

CRONOGRAMA

Clase 1. 07 de Septiembre de 2018
Clase 2. 14 de Septiembre de 2018
Clase 3. 21 de Septiembre de 2018
Clase 4. 28 de Septiembre de 2018
Clase 5. 05 de Octubre de 2018
Clase 6. 12 de Octubre de 2018
Clase 7. 26 de Octubre de 2018
Clase 8. 02 de Noviembre de 2018
Clase 9. 09 de Noviembre de 2018
Clase 10. 16 de Noviembre de 2018

BIBLIOGRAFIA GENERAL

- Anderson, C. (2008), "The end of theory: the data deluge makes scientific theory obsolete", en *Wired*, disponible en <https://www.wired.com/2008/06/pb-theory/>.
- AA. VV. (2008), "On Chris Anderson's The End of Theory", disponible en, https://www.edge.org/discourse/the_end_of_theory.html.
- Blei, D. (2012), "Probabilistic Topic Models", en *Communications of the ACM*, Vol. 55, n° 4, (Disponible en <http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf>)
- Cantú, F. (2017), "The fingerprints of fraud: evidence from Mexico's 1988 presidential election", paper presentado en *PolMeth XXXIV*, disponible en https://polmeth.polisci.wisc.edu/Papers/1988_1.pdf.
- Grimmer, J. (2018), "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together", *Political Science & Politics*, 48(1), 80-83, disponible en http://stanford.edu/~jgrimmer/bd_2.pdf.
- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2016), *An Introduction to Statistical Learning with Applications in R*, Berlín: Springer, disponible en <http://www-bcf.usc.edu/~gareth/ISL/>.
- Lamanna F., Lenormand M., Salas-Olmedo M.H., Romanillos G., Gonçalves B., Ramasco JJ. (2018), "Immigrant community integration in world cities". *PLoS ONE*, 13(3), disponible en: <https://doi.org/10.1371/journal.pone.0191612>.
- Sanz, C., Zamberlan, F., Erowid, E., Erowid, F. y Tagliazucchi, E. (2018) "The Experience Elicited by Hallucinogens Presents the Highest Similarity to Dreaming within a Large Database of Psychoactive Substance Reports", en *Frontiers in Neuroscience*, enero 2018, disponible en <https://www.frontiersin.org/articles/10.3389/fnins.2018.00007/full>.
- Sosa Escudero, Walter (2017), "Big Data y Aprendizaje Automático: Ideas y Desafíos para Economistas", versión preliminar, no publicada, disponible en <http://waltersosa.weebly.com/uploads/2/2/1/8/22189288/aaep2017bigdatawse.pdf>

- Taddy, M. (2018), “ The Technological Elements of Artificial Intelligence”, *NBER Working Paper*, n° 24301, disponible en <http://www.nber.org/papers/w24301>.
- Van der Plaas, J. Python Data Science Handbook. Essential Tools for Working with Data, O'Reilly, disponible en <https://jakevdp.github.io/PythonDataScienceHandbook/>.
- Wallach, H. (2014), “Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency”, disponible en <https://medium.com/@hannawallach/big-data-machine-learning-and-the-social-sciences-927a8e20460d>.
- Weiss, D., Nelson, A., Gibson, H. et al (2018), “A global map of traveling time to cities to asses inequalities in accesibility in 2015”, *Nature*, Vol 01, 01/2018, disponible en <https://www.nature.com/articles/nature25181>.