Ciencia de Datos

CIDE- Primavera 2021

Carlos Grandet (mailto:carlos.grandet@gmail.com)

Descripción General

En este curso se hará una presentación rigurosa de los métodos más utilizados en ciencia de datos así como sus diversas aplicaciones. Los tres grandes temas que se tratarán son: programación, datos y métodos y implementación de modelos, herramientas fundamentales en analítica.

Programación: Programar en algún lenguaje ampliamente utilizado es fundamental para poder manejar el acervo de datos que las empresas tienen. En este curso se utilizará Python, uno de los lenguajes más utilizados en computación científica y ciencia de datos y similar a Matlab o R. Se utilizarán los módulos numéricos y de manejo de bases de datos Numpy, Scipy y Pandas. Para visualización se utilizarán Matplotlib, Bokeh y Seaborn. Todos disponibles gratis con la descarga de Anaconda. Hay varios módulos para estimación que incluyen Scikit-Learn, PyBrain y StatsModel.

Datos y Métodos: el núcleo del curso es la presentaciós de los algorítmos más utilizados. Se presentarán métodos econométricos, de aprendizaje estadístico y minería de datos. Es recomendable que los estudiantes hayan tomado algún curso de econometría.

Implementación de modelos: una vez se han estimado diferentes modelos, es fundamental tener la capacidad de llevar a cabo todo el flujo de ciencia de datos para poder aplicar modelos en escenarios aplicados. Se verán técnicas principales de entrenamiento y validación de modelos.

Notas de Clase

Semanalmente iremos actualizando las notas de clase que pueden descargar directamente de Github (https://github.com/cgrandet/CIDE_DS)

Temas y fechas

Contenido	Fecha	Referencias
Introducción al curso	01/02/22	4
Python	03/02/22	19,20
Python	08/02/22	19,20
Fundamentos de ML	10/02/22	2: Chapter 3,5,6
Fundamentos de ML	15/02/22	2: Chapter 3,5,6
Modelos lineales	17/02/22	2: Chapter 4
Modelos lineales	22/02/22	2: Chapter 4
Modelos lineales	24/02/22	2: Chapter 7
Modelos no lineales	01/03/22	2: Chapter 7
Modelos no lineales	03/03/22	2: Chapter 8
Modelos no lineales	08/03/22	2: Chapter 8
Modelos no lineales	10/03/22	2: Chapter 9
Modelos no lineales	15/03/22	2: Chapter 9
Modelos no supervisados	17/03/22	2: Chapter 12
Modelos no supervisados	22/03/22	2: Chapter 12
Examen Parcial	24/03/22	
Diseño de experimentos	29/03/22	12,13
Diseño de experimentos	31/03/22	15-18
El flujo de machine learning	05/04/22	
Workshop: Limpieza de bases de datos	07/04/22	
Limpieza de bases de datos	19/04/22	
Workshop: EDA and sampling	21/04/22	
EDA	26/04/22	
Workshop: Feature Engineering	28/04/22	
Feature Engineering	03/05/22	
Workshop: Hyperparameter Tuning and model selection	05/05/22	
Hyperparameter tuning	10/05/22	
Workshop: Feature importance and partial effects	12/05/22	
Feature importance and partial effects	17/05/22	
Model Fairness	19/05/22	
Model Fairness	24/05/22	
Data y Model Drift	26/05/22	

Tareas y Evaluación

1. Cada dos semanas se asignará un trabajo práctico con aplicaciones de los métodos aprendidos.
2. La evaluación del curso será así:

    a. **Tareas**: 40%

        * Fecha: Quincenalmente

    b. **Primer parcial**: 20%

        * Fecha: 24 de marzo

    c. **Proyecto final** 30%

    d. **Presentación en clase** 10%

Proyecto

El proyecto consistirá en la implementación de un modelo de machine learning supervisado para resolver un problema a elegir. Se podrá hacer en grupos de 2 o 3.

El 31 de marzo se deberá entregar la propuesta del proyecto con lo siguiente: objetivo a cumplir, bases de datos que se usará y modelos a emplear.

Los proyectos se entregarán el 10 de junio. Se hará una sesión para presentarlos (fecha TBD). El proyecto deberá contar con las siguiente:

1) código bien documentado

2) presentación de resultados

Presentación

La presentación en clase consistirá en una exposición de métodos para el entrenamiento de modelos. Será en grupos y deberán investigar el tema a elegir dentro de los 5 posibles temas (marcados como Workshop en el temario)

Referencias

Minería de datos y aprendizaje de máquinas:

1. Hastie, Tibshirani y Friedman (2009), The Elements of Statistical Learning. Data Mining, Inference and Prediction, NY: Springer. Disponible
2. James, Witten, Hastie y Tishshirani (2021), An Introduction to Statistical Learning: with applications in R, NY: Springer online (https://hastie.su.domains/ISLRv2_website.pdf)
3. Tsiptsis y Chorianopoulos (2009), Data Mining Techniques in CRM: inside customer segmentation, Wiley
4. Varian, Hal (2014), “Big Data: New Tricks in Econometrics”, Journal of Economic Perspectives, 28 (2)
5. Wu, et.al. (2008), “Top 10 algorithms in data mining”, Knowledge and Information Systems, 14.
6. Wedel y Kamakura (2000), Market Segmentation. Conceptual and Methodological Foundations. 2nd Edition. Analytics
7. Davenport, Thomas (2006), “Competing on Analytics”, Harvard Business Review, Jan.
8. Davenport y Harris (2007), Competing on Analytics, Cambridge: Harvard Business School Publishing Coporation.
9. Davenport y Harris (2007), “The Dark Side of Customer Analytics”, HBR Case Study and Commentary, May.
10. Davenport y Patil (2012), “Data Scientist: The Sexiest Job of the 21st Century”, HBR, October
11. Loveman (2003), “Diamonds in the Data Mine”, HRB, May Diseño de Experimentos
12. Angrist y Pischke (2009), Mostly Harmless Econometrics: An Empiricist´s Companion, Princeton: Princeton University Press
13. Wooldridge (2010), Econometric Analysis of Cross Section and Panel Data
14. Greene (2011), Econometric Analysis
15. Duflo, Glennerster y Kremer, “Using Randomization in Development Economics Research: A Toolkit”.
16. Almquist y Wyner, “Boost Your Marketing ROI with Experimental Design”
17. Andersen y Simester, “A Step-by-Step Guide to Smart Business Experiments”, Harvard Business Review, Marzo 2011.
18. Khanna y New (2008), “Revolutionizing the Workplace: A Case Study of the Future of Work Program at Capital One”, Human Resource Management, 47 (4) Programación, Python y Visualización
19. McKinney, Python for Data Analysis, O’Reilly.
20. Allen Downey, Think Python. How to Think Like a Computer Scientist.
21. Cyrille Rossant, IPython Interactive Computing and Visualization Cookbook.
22. Edward Tufte, The Visual Display of Quantitative Information.
23. Willi Richert, Building Machine Learning Systems with Python Customer Lifetime Value, Credit Scoring, Visualización Dinámica
24. Murray (2013), Interactive Data Visualization for the Web, Sebastopol: O’Reilly Media
25. Bokeh: Python Library: bokeh.pydata.org
26. Gupta, et.al. “Modeling Customer Lifetime Value”, Journal of Service Research,9 (2)
27. Gupta y Lehman (2003), “Customers as Assets

In [ ]: