Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning
para Economía Aplicada
MECA 4107
Créditos: 4
2024-2

1 Información del equipo pedagógico y horario atención a estudiantes

Profesor: Ignacio Sarmiento-Barbieri (i.sarmiento@uniandes.edu.co)

- Horario Clase: Martes 6:00 p.m. - 8:50 p.m.
- Web del Curso: Bloque Neón
- Horario de atención a estudiantes: Vía Slack o hacer cita en https://calendly.com/i-sarmiento/horarios-atencion-estudiantes

Profesor Complementario: Andres Felipe Rengifo Jaramillo (a.rengifoj@uniandes.edu.co)

- Horario Clase: Viernes 6:00 p.m. - 7:20 p.m. Virtual
- Horario de atención a estudiantes: Vía Slack

2 Descripción del curso

Este es un curso con un enfoque especial en herramientas relevantes para economistas y ciencias sociales. Está destinado a estudiantes interesados en investigación aplicada y/o análisis de datos grandes y no estructurados. Problemas de predicción e inferencia, con especial énfasis en inferencia causal, atraviesan transversalmente al curso.

Mediante una combinación de contenido asincrónico, clases sincrónicas magistrales y complementarias, talleres grupales, y quices los estudiantes adquirirán las herramientas estadísticas y computacionales necesarias para responder varias preguntas en economía y en una gran cantidad de subcampos en investigación aplicada. Se hará énfasis especial en el análisis de datos reales, y la aplicación de metodologías específicas; ejemplos incluyen encuestas de hogares, precios de propiedades, datos de internet y redes sociales.

3 Resultados de aprendizaje

- Aplicar las técnicas provenientes de la ciencia de datos, las ciencias computacionales, y la estadística usando la visión de la economía para resolver problemas puntuales identificados en el contexto.
- Contrastar los distintos algoritmos y su conveniencia para contestar preguntas económicas y sociales con base en criterios relacionados con la naturaleza de problemas económicos y sociales.
- Implementar procesos técnicos para el manejo cuantitativo de datos que surgen de distintas fuentes: páginas web, encuestas, geoespaciales, texto, etc., para resolver problemas económicos y sociales.

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning
para Economía Aplicada
MECA 4107
Créditos: 4
2024-2

- Generar conclusiones y recomendaciones sobre preguntas relevantes a las ciencias sociales por medio del manejo, análisis y síntesis de bases de datos con gran número de observaciones y variables.
- Aplicar el software `R` y su ecosistema para análisis estadístico, de big data y machine learning.

4 Cronograma

Table 1: Cronograma *Tentativo*

Semana	Tema	¿Qué estudiaremos?
1	Introducción y Motivación. Regresión Lineal	Esta semana nos preguntaremos qué es el Big Data y el Machine Learning. Exploraremos la distinción entre predicción e inferencia causal, y las consideraciones al aplicar modelos de aprendizaje de máquinas, incluyendo aspectos éticos. Estudiaremos la regresión lineal como el algoritmo básico del Aprendizaje de Máquinas, haciendo énfasis en las similitudes y diferencias con la econometría, discutiendo la diferencia entre estimar y predecir.
2	Detalles Regresión Lineal	En la segunda semana estudiaremos cómo se utilizan las ecuaciones normales en la regresión lineal y su importancia computacional. Aprenderemos sobre el gradiente descendente y cómo se aplica en la regresión lineal para optimizar los parámetros del modelo. Estudiaremos el Teorema de Frisch Waugh Lovell y su importancia en la práctica, tanto desde el punto de vista computacional como para el diagnóstico de modelos lineales.
3	Incertidumbre: Bootstrap y técnicas de remuestreo	Esta semana exploraremos la incertidumbre centrándonos en el Bootstrap, una técnica computacional que nos permite cuantificar la incertidumbre asociada a los estimadores o los métodos de aprendizaje de máquinas.
4	Generalización. Sobreajuste. Evaluación predictiva fuera de muestra	En la cuarta semana discutiremos herramientas para analizar y entender la generalización de los modelos de Machine Learning, es decir, su desempeño fuera de muestra. Estudiaremos el concepto de sobreajuste y el dilema sesgo-varianza, así como las formas de evaluar el error fuera de muestra tanto de una perspectiva clásica como moderna a través de la validación cruzada.
5	Selección de Modelos y Regularización. Ridge	Esta semana exploraremos la selección de modelos desde una perspectiva clásica y moderna. Desde la perspectiva clásica, estudiaremos el *subset selection*, y desde la perspectiva moderna, los métodos de regularización. Específicamente, nos centraremos en el estimador de Ridge. Analizaremos tanto la teoría detrás de este método como su implementación práctica.

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning para Economía Aplicada
MECA 4107
Créditos: 4
2024-2

Table 2: Cronograma *Tentativo* (Cont.)

6	Regularización: LASSO y Elastic Net	En la sexta semana continuamos con los métodos de regularización centrándonos en LASSO y Elastic Net. Estudiaremos cómo estos métodos ayudan a la selección de modelo a través de la penalización. Analizaremos tanto la teoría como la implementación práctica. *Tentativamente*, si el tiempo lo permite, exploraremos su uso en inferencia causal.
7	Riesgo, Probabilidad y Clasificación.	Esta semana nos enfocaremos en la clasificación, comenzaremos entendiendo el vínculo Riesgo-Probabilidad-Clasificación. Nos centraremos en el modelo Logit para estimar probabilidades y los métodos numéricos, como el gradiente descendente, para resolverlo. Luego nos moveremos a otros tipos de modelos que nos permiten estimar probabilidades.
8	Métricas de desempeño en clasificación. Desbalance de clases	En la octava semana continuamos estudiando clasificación, centrándonos en las métricas de desempeño de estos modelos: precisión, recall, F-score y AUC-ROC, y cómo se interpretan en el contexto de la clasificación. Discutiremos qué es el desbalance de clases y por qué es un problema en los modelos de clasificación, especialmente en problemas económicos. Exploraremos técnicas comunes para manejar el desbalance de clases y cómo aplicar estas técnicas para mejorar el desempeño predictivo de los modelos.
9	Árboles, Bagging y Bosques Aleatorios	Esta semana nos enfocaremos en métodos basados en árboles de regresión y clasificación. Estudiaremos qué son los árboles y sus ventajas y limitaciones, especialmente en la captura de relaciones no lineales y su carácter no paramétrico. Exploraremos el método de Bagging y Bosques Aleatorios como una extensión para mejorar la estabilidad, precisión y desempeño predictiva de los modelos. Discutiremos sobre la interpretabilidad de los modelos. *Tentativamente*, si el tiempo lo permite, exploraremos su uso en inferencia causal.
10	Boosting	En la décima semana nos enfocaremos en el método de Boosting. Estudiaremos qué es Boosting explorando las principales técnicas como AdaBoost, Gradient Boosting y XGBoost, y cómo estas mejoran la precisión de los modelos al enfocarse en los errores de predicción. Analizaremos las ventajas que ofrece Boosting en comparación con otros métodos, así como sus posibles desventajas. Discutiremos cómo se interpretan y evalúan los resultados y en qué situaciones son especialmente útiles para capturar relaciones complejas y no lineales en los datos.
11	Datos espaciales, validación cruzada espacial	Esta semana nos enfocaremos en los datos espaciales y la validación cruzada espacial. Estudiaremos qué son los datos espaciales y analizaremos las características y desafíos específicos del análisis de datos espaciales, así como las técnicas para manejarlos y analizarlos en distintas aplicaciones. Exploraremos qué es la validación cruzada espacial y cómo se diferencia de la validación cruzada tradicional. Veremos cómo implementar la validación cruzada espacial para garantizar que los modelos sean robustos y generalizables en el contexto de datos espaciales.

Table 3: Cronograma *Tentativo* (Cont.)

12	Redes neuronales de una sola capa	En la decimosegunda semana del curso iniciaremos el estudio del aprendizaje profundo y las redes neuronales. Nos enfocaremos en las redes de una sola capa, también conocidas como perceptrones. Analizaremos los componentes básicos de un perceptrón y el proceso de entrenamiento enfocándonos en el algoritmo de backpropagation. Aprenderemos cómo ajustar los hiperparámetros y mejorar el desempeño de estos modelos en el análisis económico.
13	Redes neuronales profundas	Esta semana continuamos con redes neuronales enfocándonos en las redes neuronales profundas y cómo se diferencian de las redes neuronales de una sola capa. Exploraremos técnicas de regularización como dropout y batch normalization, y cómo estas ayudan a mejorar el rendimiento de las redes profundas. Discutiremos en qué aplicaciones económicas pueden ser especialmente útiles las redes neuronales profundas y las consideraciones que deben tenerse en cuenta al evaluar y ajustar los hiperparámetros para garantizar un rendimiento óptimo. *Tentativamente* , si el tiempo lo permite, estudiaremos las redes neuronales LSTM (Long Short-Term Memory) y su aplicación en el análisis de series temporales.
14	Super Learners	Esta semana exploraremos el concepto de Super Learners. Estudiaremos cómo se combinan múltiples modelos para crear un Super Learner que optimiza el rendimiento predictivo y las ventajas de utilizar Super Learners en comparación con modelos individuales. Analizaremos las técnicas para entrenar y validar un Super Learner, como la validación cruzada y la ponderación de modelos. Discutiremos qué consideraciones deben tenerse en cuenta para asegurar que los Super Learners sean robustos y generalizables.
15	Texto como datos y reducción de dimensión	En la última semana del curso, nos enfocaremos en el uso del texto como datos y la reducción de dimensionalidad. Exploraremos técnicas básicas para el procesamiento del lenguaje natural (NLP). Discutiremos la reducción de dimensión, como el PCA y LDA, y su importancia en el análisis de datos a gran escala, especialmente en el contexto del texto.

5 Referencias

- Davidson, R., & MacKinnon, J. G. (2004). Econometric theory and methods
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning (ISLR)
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
- Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied

causal inference powered by ML and AI. arXiv preprint arXiv:2403.02467.

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

6 Metodología

La metodología del curso combina clases magistrales presenciales y complementarias virtuales, cuestionarios y talleres.

El objetivo de la clase magistral es cubrir los principales conceptos, modelos y metodologías del aprendizaje de máquinas de forma teórica y a través de aplicaciones prácticas que ilustran los conceptos. La presentación se hará a través de diapositivas, cuadernos interactivos, tablero y marcador.

Las clases complementarias servirán de apoyo a las magistrales y en el acompañamiento de la realización de los talleres. En las clases complementarias se desarrollarán aplicaciones y se resolverán inquietudes específicas respecto al material o los talleres.

Se espera que los estudiantes realicen las actividades asincrónicas respectivas antes de cada clase magistral, estudien las lecturas disponibles en la página del curso y repliquen las aplicaciones presentadas por el profesor. Todas las actividades asincrónicas están pensadas para facilitar el aprendizaje y posterior desarrollo de las actividades individuales y grupales.

Para un desarrollo exitoso del curso se espera que los estudiantes asistan a todas las clases, participen en clase y a través de los canales de Slack.

La evaluación del aprendizaje se realizará a través de cuestionarios individuales y tres talleres prácticos grupales que servirán para que los estudiantes pongan a prueba los conocimientos adquiridos.

7 Evaluaciones

- Quices. Los estudiantes tendrán 8 quices individuales al finalizar los módulos en Bloque Neón que evaluarán el aprendizaje individual.

- Talleres. Los estudiantes realizarán trabajos prácticos grupales donde los grupos no podrán superar los 4 miembros. Habrá 3 talleres durante el cursado. Los talleres serán entregados vía Bloque Neón y deberán contar con un repositorio en GitHub. Se espera que todos los miembros hagan contribuciones al repositorio del taller. La calificación del taller se verá reducida si no hay evidencia de contribución de todos los miembros.

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning
para Economía Aplicada
MECA 4107
Créditos: 4
2024-2

Table 4: Puntajes

	Puntaje Individual	Puntaje Total	Fecha entrega
Quices		40%	
Quiz 0*	0%		Agosto 11, 2024
Quiz 1	5%		Agosto 18, 2024
Quiz 2	5%		Septiembre 1, 2024
Quiz 3	5%		Septiembre 15, 2024
Quiz 4	5%		Septiembre 29, 2024
Quiz 5	5%		Octubre 20, 2024
Quiz 6	5%		Noviembre 3, 2024
Quiz 7	5%		Noviembre 17, 2024
Quiz 8	5%		Diciembre 1, 2024
Talleres		60%	
Taller 1	20%		Septiembre 15, 2024
Taller 2	20%		Octubre 20, 2024
Taller 3	20%		Diciembre 1, 2024

Nota: * *Opcional*

Nota bene: El objetivo de permitir el trabajo en equipos es fomentar la discusión y colaboración en el proceso de aprendizaje, no simplemente dividir las tareas. Se espera que cada miembro del equipo contribuya de manera significativa en cada parte de cada taller. Además, cada estudiante es responsable de todo el contenido del taller, independientemente de cómo se organicen para trabajar en equipo.

Sistema de aproximación de notas definitiva

Las calificaciones definitivas de las materias serán numéricas de uno punto cinco (1.50) a cinco (5.00), en unidades, décimas y centésimas. La calificación aprobatoria mínima será de tres (3.0). En este curso se aproximará la nota a la centésima más cercana. Por ejemplo, si el cálculo del cómputo es 3.245, la nota final se aproximará a 3.25; si el resultado del cálculo es 2.994 la nota final será de 2.99

8 Asistencia

Se espera que los estudiantes asistan a todas las clases. Si los estudiantes faltan a más del 20% sin excusa válida de las clases presenciales se penalizará hasta un 10% la nota final del curso.

Universidad de los Andes
Facultad de Economía

Big Data and Machine Learning
para Economía Aplicada
MECA 4107
Créditos: 4
2024-2

9 Políticas generales de los cursos de Economía y fechas importantes

Los estudiantes deben consultar este enlace, donde se encuentran las reglas sobre asistencia a clase, excusas válidas, fraude académico y faltas disciplinarias, reclamos, políticas de bienestar y fechas importantes del semestre.