Universidad Tecnológica de la Mixteca

Clave DGP: 200089

Doctorado en Modelación Matemática

00012

PROGRAMA DE ESTUDIOS

NOMBRE DE LA ASIGNATURA
Modelación avanzada para ciencia de datos

SEMESTRE	CLAVE DE LA ASIGNATURA	TOTAL DE HORAS
Segundo	**292201**	**80**

OBJETIVO(S) GENERAL(ES) DE LA ASIGNATURA

Estudiar, analizar y comprender a profundidad los modelos para manejo de datos faltantes, detección de atípicos, de transformación de variables y de aprendizaje máquina, con el objetivo que el estudiante sea capaz de ajustarlos o diseñar nuevos para resolver algunos de los problemas abiertos que se presentan en ciencia de datos.

TEMAS Y SUBTEMAS

1. Algunos problemas comunes en ciencia de datos

1.1. La maldición de la dimensión.
1.2. Manejo de datos perdidos y faltantes (Congenialidad de la imputación).
1.3. Problemas en selección de características.
1.4. Efectos de los datos atípicos.
1.5. Efectos de la similaridad y métricas.
1.6. El problema de la descomposición matricial en matrices de rango bajo.
1.7. Problemas abiertos en reducción de la dimensionalidad espectral.
1.8. Problemas comunes en validación de modelos de aprendizaje automático.
1.9. Problema de elección del k óptimo en el paradigma de vecinos más cercanos.

2. Modelos para manejo de datos faltantes

2.1. Eliminación de datos: modelos supervisados y no supervisados.
2.2. Eliminación Hot-Deck
2.3. Imputación Bayesiana.
2.4. Imputación múltiple propia.
2.5. Imputación para datos continuos basados en modelos normales lineales.
2.6. Imputación para datos no continuos basados en modelos generalizados lineales.
2.7. Modelos de suavizado.
2.8. Modelo Expectation-Maximization.

3. Modelos para detección de atípicos

3.1. Modelo ROCK
3.2. Squeezer
3.3. k-ANMI
3.4. k-modes
3.5. Greedy
3.6. AVF
3.7. Road
3.8. mRMR
3.9. NMIFS
3.10. Selección de variables no supervisado

4. Modelos para transformación de variables

4.1. Modelos de codificación.
  4.1.1. Codificación binaria.
  4.1.2. One hot encoding.
  4.1.3. Codificación entera.
  4.1.4. Para texto: Flatten and filter, estandarización y normalización l^2
  4.1.1. Para data stream.
4.2. Modelos de discretización.
  4.2.1. Bining.
  4.2.2. Cluster Analysis
  4.2.2. Mediante árboles de decisión.
  4.2.2. Mediante análisis de correlación.

PROGRAMA DE ESTUDIOS

5. Modelos de aprendizaje automático
- 5.1. Modelos generalizados lineales.
- 5.2. Modelos lineales mixtos.
- 5.2. Modelos marginales basados en verosimilitud.
- 5.3. Modelos condicionales.
- 5.4. Modelo generalizado lineal mixto GLMM.
- 5.5. Modelos de máquina de soporte vectorial.
- 5.6. Modelos basados en kernel.

ACTIVIDADES DE APRENDIZAJE

Sesiones dirigidas por parte del profesor en la que se presentan los conceptos poniendo énfasis en los fundamentos matemáticos de cada modelo. Se sugiere utilizar algún Notebook como Collaboratory o Jupyter o Visual Studio Code para realizar programas con el lenguaje Python, así como Kaggle y GitHub para compartir y descargar algoritmos programables. Se recomienda ampliamente impartir el curso en un laboratorio con equipo de cómputo disponible para cada estudiante. El contenido será abordado a profundidad y realizada la modelación con fines de práctica de cada uno de los modelos en cada unidad.

CRITERIOS Y PROCEDIMIENTOS DE EVALUACIÓN Y ACREDITACIÓN

Se aplican por lo menos tres exámenes parciales cuyo promedio equivale al 50% de la calificación final, el 50% restante se obtiene de un examen final. Otras actividades que se consideran para la evaluación son las participaciones en clase, asistencias a clases y el cumplimiento de tareas.

BIBLIOGRAFÍA

Básica:

1. Yulei He, Guangyu Zhang y Chiu-Hsieh Hsu. Multiple imputation of missing data in practice, Basic theory and Analysis Strategies, CRC Press, Taylor & Francis Group.
2. N. N. R. Ranga Suri, Narasimha Murty M y G. Athithan. Outlier detection: Techniques and applications, A data mining perspective, Springer.
3. P. E. McKnight. K. M. McKnight, S. Sidani y A. J. Figueredo. Missing Data, A gentle introduction, The Guilford Press.
4. C. Campbell y Yiming Yin. Learning with support vector machines, Synthesis Lectures on artificial intelligence and machine learning No. 10, Morgan & Claypool publishers.

Consulta:

1. H. Strange y R. Zwiggelaar. Open problems in spectral dimensionality reduction, Springer.
2. J. Leskovec, A. Rajaraman y J. D. Ullman. Mining of Massive Datasets, Cambridge University Press.
3. B. Ratner, Statistical and Machine Learning Data Mining, Techniques for Better Predictive Modeling and Analysis of Big Data, second edition, CRC Press, Taylor & Francis Group.
4. M. Bahri , A. Bifet , S. Maniu y H. Murilo Gomes, Survey on Feature Transformation Techniques for Data Streams, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20).

PERFIL PROFESIONAL DEL DOCENTE

Estudios de Doctorado en Matemáticas o en Estadística.

Vo.Bo
DR. JOSÉ ANIBAL ARIAS AGUILAR
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO

DIVISION DE ESTUDIOS DE POSGRADO

AUTORIZÓ
DR. RAFAEL MARTÍNEZ MARTÍNEZ
VICE-RECTOR ACADÉMICO

VICE-RECTORIA ACADÉMICA