UNIVERSIDAD DEL VALLE DE GUATEMALA
CAMPUS CENTRAL
FACULTAD DE CIENCIAS Y HUMANIDADES

UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Iniciativa Académica de Aprendizaje Estadístico

1 Identificación

Curso:	MM3024 – Seminario 2 de Matemática	**Créditos:**	4
Ciclo:	Primero	**Requisitos:**	Data Mining, Data Science
Año:	2024		Estadística Matemática
			Métodos Numéricos 2
Profesor:	Alan Reyes–Figueroa	**Horario:**	Lunes y miércoles – 19:50-21:25
Email:	agreyes	**Sala:**	CIT-517.

Sitio Web del Curso:

- https://pfafner.github.io/sl2024

Office Hours:

- Por solicitud del estudiante, o pueden enviar sus dudas por correo electrónico.

2 Descripción

Este es un curso introductorio al aprendizaje estadístico, con énfasis principalmente en los fundamentos matemáticos y estadísticos de los principales algoritmos de aprendizaje automático y reconocimiento de patrones. El tema central del curso es el estudio de métodos para obtener información útil a partir de datos. Abordamos temas principales como el aprendizaje supervisado y no supervisado, los modelos de regresión, y algunos tópicos recientes como el aprendizaje profundo. Al final del curso, los estudiantes comprenderán los fundamentos de los algoritmos más populares del aprendizaje estadístico. Este es un curso integrador, donde se unen los conocimientos adquiridos a través de los cursos de la carrera de matemáticas, y herramientas de computación. Se requiere que el estudiante tenga un conocimiento de diversas áreas de matemática, estadística y que domine al menos un lenguaje de programación.

El curso inicia con una introducción a los métodos para reducción de dimensión como: componentes principales (PCA), componentes independientes, *auto-encoders* y variables latentes, así como algoritmos de *manifold learning*, los cuales servirán para obtener representaciones adecuadas y visualizaciones de datos. En seguida, se estudian métodos para clasificación no supervisada como: el algoritmo K-*medias* y sus variantes, mezclas gaussianas, agrupamiento jerárquico y agrupamiento espectral, así como otros algoritmos de carácter más geométrico. Haremos también un recorrido por los métodos más comunes para clasificación supervisada, como: el clasificador bayesiano óptimo, el clasificador KNN (K-*nearest neighbors*), el perceptrón y la regresión logística, las máquinas de vectores de soporte y los árboles de decisión, para continuar con algunos métodos de Seguidamente se estudiarán modelos estadísticos de regresión lineal y no-lineal. Finalmente el curso se completa con una introducción a las redes neuronales artificiales. En todos los temas se hará énfasis en los fundamentos matemáticos de cada algoritmo. El curso asume el conocimiento de conceptos estadísticos básicos, como variables aleatorias, distribuciones, independencia, covarianza y correlación, entropía. Cuando sea conveniente, se hará un repaso de estos conceptos.
El curso cuenta con una parte práctica extensiva, en la que el estudiante implementará en código computacional cada uno de los algoritmos estudiados. Parte fundamental del curso es utilizar las herramientas aprendidas en varios proyectos aplicados donde se trabajará con datos reales provenientes de diversas áreas: datos socio-económicos, datos de movilidad, datos médicos, imágenes, datos financieros, e ilustrar los resultados mediante informes y seminarios.

3 Competencias a Desarrollar

Competencias genéricas

1. Piensa de forma crítica y analítica.
2. Resuelve problemas de forma estructurada y efectiva.
3. Desarrolla habilidades de investigación y habilidades de comunicación a través de seminarios y presentaciones ante sus colegas.

Competencias específicas

1.1 Entiende y domina los fundamentos matemáticos que formaliza los algoritmos principales en la ciencia de datos y el aprendizaje estadístico.

1.2 Conoce y domina los principales métodos de clasificación y predicción de datos.

1.3 Comprende los conceptos estadísticos subyacentes a los modelos de regresión de datos univariados y multivariados.

2.1 Aplica métodos y técnicas para la exploración de datos multivariados de forma efectiva. Aplica técnicas de reducción de dimensionalidad, cuando sea conveniente.

2.2 Aplica de forma efectiva técnicas de visualización de datos, para comunicar resultados sin ambigüedad o desinformación.

2.3 Utiliza un enfoque global para resolver problemas. Utiliza herramientas auxiliares en su solución, como distribuciones, inferencia estadística, optimización, algoritmos de aprendizaje automático.

3.1 Desarrolla todas las etapas de una investigación o proyecto aplicado donde se utilizan elementos del análisis de datos: anteproyecto, exploración de datos, diseño experimental, metodología, predicción y conclusiones.

3.2 Escribe un reporte técnico sobre la solución de un problema en análisis de datos, usando datos reales. Concreta un análisis riguroso y conclusiones importantes.

3.3 Comunica de manera efectiva, en forma escrita, oral y visual, los resultados de su investigación.

4 Metodología Enseñanza Aprendizaje

El curso se desarrollará durante diecinueve semanas, con cuatro períodos semanales de cuarenta y cinco minutos para desenvolvimiento de la teoría, la resolución de ejemplos y problemas, comunicación didáctica y discusión. Se promoverá el trabajo colaborativo de los estudiantes por medio de listas de ejercicios y proyectos.

Durante el curso se promoverá la revisión bibliográfica y el auto aprendizaje a través de la solución de los ejercicios indicados, así como el desarrollo de proyectos aplicados. Se espera que el estudiante desarrolle su trabajo en grupo o individualmente, y que participe activamente y en forma colaborativa durante todo el curso.

5 Contenido

1. Repaso de conceptos estadísticos: Variables aleatorias discretas y continuas. Distribuciones. Valor esperado. Varianza. Entropía. Covarianza y correlación. Introducción a la inferencia estadística. El método de máxima verosimilitud. Funciones de pérdida, *score* e información.

2. Métodos exploratorios para datos multivariados: Visualización y resumen de la dependencia entre variables. Métodos de proyección: Descomposición en valores singulares (SVD). Análisis de componentes principales (PCA). Re-escalamiento multidimensional. Kernel PCA. Análisis de componentes independientes (ICA). Reducción de la dimensionalidad: Factoración de matrices no-negativas (NNMF). Variables latentes. Otros tópicos: El modelo de Kohonen. *Manifold learning*: Isomap, Local Linear Embbedding, Spectral Embeding. SOM. Funciones *kernel* y estimación empírica de distribuciones.

3. Aprendizaje no-supervisado: Métodos de agrupamiento. Métodos geométricos vs. métodos probabilísticos. Métodos de agrupamiento jerárquico. Métodos locales: $k-$medias, $k-$medianas, $k-$medoides. Dendrogramas. Algoritmos basados en mezclas y densidades. Algoritmo EM. Agrupamiento espectral. Métricas para evaluar modelos.

4. Aprendizaje supervisado: El clasificador bayesiano. Análisis discriminante. $k-$*nearest neighbors*. Regresión logística. Máquinas de soporte vectorial (SVM). Métodos *kernel*. Árboles de Decisión. Modelos *ensemble*. Random forests. *Bagging*, *Boosting*, *Stacking*. Redes neuronales artificiales. *Auto-encoders*. Validación cruzada y selección de modelos.

5. Modelación estadística y predicción: Mínimos cuadrados. Modelos de regresión lineal (generalizada). Pruebas de hipótesis y gráficos de diagnóstico. Selección de variables. Métodos de regularización: Ridge (L_2), LASSO (L_1), *Elastic-net* (L_0). Criterios de selección de modelos: AIC, BIC. Mínimos cuadrados parciales.

6 Bibliografía

Textos:

- G. Strang (2019). *Linear Algebra and Learning from Data*. Cambridge Press.
- R. Duda, P. Hart, D. Stork (2000). *Pattern classification*. Wiley.

Referencias adicionales

- C. Bishop (2000). *Pattern Recognition and Machine Learning*. Springer
- T. Hastie, R. Tibshirani, J. Friedman (2013). *The Elements of Statistical Learning*. Springer.
- K. Murphy (2012). *Machine Learning: a Probabilistic Perspective*. MIT Press.
- G. James, D. Witten, T. Hastie, R. Tibshirani (2008). *An Introduction to Statistical Learning with Applications in R*. Springer.
- A. Izenman (2008). *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*. Springer.
- K. Fukunaga (1990). *Introduction to Statistical Pattern Recognition*. Academic Press.
- C. Giraud (2015). *Introduction to High-Dimensional Statistics*. CRC/Chapman and Hall.
- L. Devroye, L. Györfi, G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.
- S. Shalev-Shwartz, S. Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge U. Press. https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding\-machine-learning-theory-algorithms.pdf
- P. Rigollet (2015). *Mathematics for Machine Learning*. https://ocw.mit.edu/courses/mathematics/18-657-mathematics-of-machine-learning-fall-2015/lecture-notes/MIT18_657F15_LecNote.pdf.

7 Actividades de evaluación

Actividad	Cantidad aproximada	Porcentaje
Listas de ejercicios	6 a 8	55%
Proyectos	2	45%

8 Cronograma

Semana	Tópico	Fecha	Actividades
1	Introducción y motivación al curso. Probabilidad. Probabilidad condicional. Variables aleatorias.	08-12 enero	
2	Varianza, covarianza y correlación. Entropía. Distribuciones, estadísticos y resúmenes. Dependencia entre variables.	15-19 enero	
3	SVD y PCA. Intrepretación del PCA. Ejemplos y aplicaciones.	22-26 enero	
4	Variantes de PCA. Re-escalamiento multidimensional. ICA. Factoración de matrices no-negativas.	29 enero-02 febrero	
5	Otros métodos de variables latentes. Funciones kernel. Distribuciones empíricas.	05-09 febrero	
6	Métodos locales: U-map, Isomap, t-SNE, *Local Embeddings*. *Manifold Learning*, SOM.	12-16 febrero	
7	Métodos de agrupamiento jerárquico. Dendrogramas. $k-$medias, $k-$medianas, $k-$medoides.	19-23 febrero	
8	Agrupamiento espectral: vector de Fiedler, NCuts. Mezclas gaussianas. El Algoritmo EM.	26 febrero-01 marzo	
9	Métodos basados en densidad: *Mean-shift*. Otros métodos de agrupamiento.	04-08 marzo	
10	Métricas para métodos de agrupamiento. El método de $K-$vecinos más cercanos (KNN).	11-15 marzo	
11	Presentaciones del primer proyecto.	18-22 marzo	Proyecto 1
	Semana Santa	25-29 marzo	
12	El clasificador bayesiano óptimo. Ejemplos. Clasificador *Naïve Bayes*. Cotas de Error.	01-05 abril	
13	Análisis discriminante (LDA).	08-12 abril	
14	Clasificadores lineales: el clasificador logístico. El Perceptrón. Máquinas de vectores de soporte (SVM).	15-19 abril	
15	Árboles de decisión. Entropía e impureza. *Random forests*. Modelos ensamblados: *Bagging*, *Boosting*, *Stacking*.	22-26 abril	
16	Redes neuronales artificiales. Métricas para métodos de clasificación.	29 abril-03 mayo	
17	El modelo de regresión lineal ordinaria (OLS). Estimación de parámetros en regresión. Métodos de regularización.	06-10 mayo	
18	Puntos palanca, sensibilidad. Gráficos de diagnóstico. Regresión no-paramétrica.	13-17 mayo	
19	Selección de variables y modelos. Criterios de información. Métricas para clasificación. Validación cruzada.	20-24 mayo	
20	Presentación de proyectos.	27-31 mayo	Proyecto 2