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

UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Iniciativa Académica de Introducción a la Ciencia de Datos

1 Identificación

Curso:	MM3024 – Seminario 2 de Matemática	**Créditos:**	4
Ciclo:	Primero	**Requisitos:**	Álgebra Lineal 2
Año:	2022		Estadística Matemática 1
			Programación y Algoritmos
Profesor:	Alan Reyes–Figueroa	**Horario:**	Lunes y miércoles – 17:20-18:55
Email:	agreyes	**Lab:**	A definir

Sitio Web del Curso:

- https://pfafner.github.io/cd2022

Office Hours:

- Viernes de 18:00 a 20:00 hrs, o por solicitud del estudiante. También pueden enviar sus dudas por correo electrónico.

2 Descripción

Este es un curso introductorio al análisis de datos. El curso hace una introducción a los tópicos de aprendizaje de máquina y reconocimiento estadístico de patrones, ambos considerados temas de importancia actual. El curso pretende que el estudiante adquiera habilidades para el análisis de información y toma de decisiones, a partir del análisis estadístico de datos. En el curso se desarrolla una colección de herramientas matemáticas, estadísticas y computacionales, abarcando los principales temas del análisis de datos, como: clasificación supervisada, no supervisada y regresión. 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 de clasificación estadística como: el clasificador bayesiano, el clasificador Knn (k-*nearest neighbors*), la máquina de vectores de soporte y los árboles de decisión, para continuar con algunos métodos de clasficiación no-supervisada: k-*means*, agrupamiento jerárquico y agrupamiento espectral. Seguidamente se estudiará el modelo de regresión logística, y se hace un estudio de los modelos estadisticos de regresión lineal y no-lineal. Se hará una introducción a los métodos de reducción de dimensión como: componentes principales (PCA), *auto-endocers* y variables latentes. Posteriormente se estudian métodos de modelación estadística y predicción de datos. Finalmente el curso se completa com 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 de patrones.

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 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, métodología, 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. El curso cuenta con una sesión de laboratorio semanal para la implementación de algoritmos y la práctica de las técnicas de análisis de datos.

El resto del curso promoverá la revisión bibliográfica y el auto aprendizaje a través de la solución de los ejercicios del texto, y problemas adicionales, y el desarrollo de una monografía. Se espera que el alumno 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* y *Boosting*. 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:

- R. Duda, P. Hart, D. Stork (2000). *Pattern classification*. Wiley.
- 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.

Artículos:

- A. S. Bandeira (2016). *Ten Lectures and Forty-Two Open Problems in the Mathematics of Data Science*. https://people.math.ethz.ch/~abandeira/TenLecturesFortyTwoProblems.pdf

Referencias adicionales

- 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	50%
Proyectos	2	50%

8 Cronograma

Semana	Tópico	Fecha	Actividades
1	Introducción y motivación al curso. Repaso de conceptos estadísticos: Probabilidad.	10-14 enero	
2	Probabilidad condicional. Variables aleatorias. Varianza. Entropía. Covarianza y correlación.	17-21 enero	
3	Introducción a la inferencia estadística: Distribuciones, estadísticos y resúmenes.	24-28 enero	
4	Métodos exploratorios. Visualización y dependencia entre variables. SVD. PCA.	31 enero-04 febrero	
5	Interpretación de PCA. Ejemplos y aplicaciones. Variantes de PCA. Re-escalamiento multidimensional.	07-11 febrero	
6	ICA. Factoración de matrices no-negativas. Funciones kernel. Distribuciones empíricas.	14-18 febrero	
7	Métodos locales: Isomap, t-SNE, *Local Embedding*, *Manifold Learning*, SOM. *Spectral Embedding*.	21-25 febrero	
8	Métodos de agrupamiento jerárquico. Dendrogramas. $K-$medias, $K-$medianas, $K-$medoides.	28 febrero-04 marzo	
9	Mezclas gaussianas. El Algoritmo EM. Agrupamiento espectral: vector de Fiedler, NCuts.	07-11 marzo	
10	Métodos basados en densidades: Mean-shift. Métricas para métodos de agrupamiento.	14-18 marzo	
11	Modelación predictiva. El método de $K-$vecinos más cercanos.	21-25 marzo	
12	El clasificador bayesiano óptimo. Ejemplos. Clasificador *Naive Bayes*.	28 marzo-01 abril	
13	Seminario de proyectos aplicados.	04-08 abril	Proyecto 1
	Semana Santa	11-15 abril	
14	Análisis discriminante (LDA). Clasificadores lineales: el clasificador logístico.	18-22 abril	
15	El Perceptrón. Máquinas de vectores de soporte (SVM). Árboles de decisión. Entropía e impureza. *Random forests*.	25-29 abril	
16	Modelos ensamblados: *Bagging*, *Boosting*, *Stacking*. El modelo de regresión lineal ordinaria (OLS).	02-06 mayo	
17	Gráficos de diagnóstico. Otros métodos de regresión.	09-13 mayo	
18	Selección de variables y modelos: AIC, BIC. Métricas para clasificación. Validación cruzada.	16-20 mayo	
19	Introducción a las redes neuronales artificiales. Redes neuronales convolucionales.	23-27 mayo	
20	Seminario de proyectos aplicados.	30 mayo-03 junio	Proyecto 2