UNIVERSIDAD FRANCISCO MARROQUÍN
FACULTAD DE CIENCIAS ECONÓMICAS
DATA SCIENCE

UFM UNIVERSIDAD FRANCISCO MARROQUÍN VERITAS • LIBERTAS • IVSTITIA | FACULTAD de CIENCIAS ECONÓMICAS

Programa del curso Elements of Machine Learning

1 Identificación

Curso:	CS-DS002 – Elements of Machine Learning	**Créditos:**	3
Semestre:	Primero	**Requisitos:**	Álgebra Lineal
Año:	2023		Cálculo, Estadística
			Programación en Python
Profesor:	Alan Reyes–Figueroa	**Horario:**	Martes y jueves – 16:00-17:20
Email:	agreyes	**Sala:**	D-409

Sitio Web del Curso:

- https://pfafner.github.io/ml2023

Office Hours:

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

2 Descripción

Este curso es una introducción al aprendizaje estadístico de datos. El aprendizaje estadístico (*Statistical Learning*) busca encontrar distintas dependencias funcionales entre las variables independientes (estimadores) y las dependientes (respuestas). Dependiendo de la aplicación, inferir la relación entre dichas variables, o detectar grupos y patrones dentro de un conjunto de datos, o se busca predecir futuros comportamientos.

Los datos en el mundo real son complejos y contienen bastante ruido, por lo que, se han desarrollado distintas técnicas para tratar de comprenderlos. En este curso aboramos los fundamentos matemáticos y estadísticos detrás de algunos de los algoritmos más populares utilizados hoy en día y veremos usos prácticos de ellos. Nos enfocaremos principalmente en técnicas de aprendizaje no supervisado (como por ejemplo, PCA, métodos de reducción de la dimensionalidad y métodos de clustering), así como aprendizaje supervisado (regresión lineal, regresión logística, k-vecinos más cercanos, entre otros). Al final del curso haremos una introducción a las redes neuronales, y otros métodos de importancia actual.

En todos los temas buscaremos entender 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 los estudiantes implementarán 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, e ilustrar los resultados mediante informes y seminarios.

3 Competencias a Desarrollar

Objetivos generales

Transferir los fundamentos teóricos matemáticos de los algoritmos de aprendizaje estadístico para que el estudiante desarrolle el criterio necesario para saber diferenciar cuándo aplicar cada algoritmo, así como que tenga la base suficiente para leer y entender publicaciones científicas.

Objetivos específicos Al final del curso, se busca que el estudiante pueda:

1. Realizar la correcta aplicación de las técnicas utilizadas en aprendizaje estadístico y ciencia de datos.
2. Planificar y ejecutar un análisis aplicado a distintos problemas reales utilizando las técnicas y algoritmos vistos en el curso.
3. Saber diferenciar cuándo aplicar las técnicas de aprendizaje supervisado y cuándo usar las de aprendizaje no supervisado.
4. Interpretar correctamente los resultados obtenidos, de manera efectiva, y en su contexto, sin importar la jerga utilizada.
5. Poder analizar y generar crítica constructiva sobre afirmaciones realizadas en datos y evaluar decisiones basadas en los mismos.
6. Desarrollar habilidades de investigación y de comunicación a través de seminarios y presentaciones ante sus colegas.

4 Metodología Enseñanza Aprendizaje

El curso se desarrollará durante diecisiete semanas, con cuatro períodos semanales de cuarenta 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 algunas sesiones de laboratorio 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 ejercicios y problemas prácticos, laboratorios y el desarrollo de varios proyectos. 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. *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. KS y otras métricas de evaluación.

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.

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.

Blogs:

- Machine Learning Mastery J. Brownlee.
- Towards IA - Machine Learning.
- Data Science Dojo.
- Machine Learning - Carnegie Mellon University.
- BAIR -Berkeley Artificial Intelligence Research.

7 Actividades de evaluación

Actividad	Cantidad aproximada	Porcentaje
Labs y Listas de ejercicios	8	50%
Proyectos	2	50%

8 Cronograma

Semana	Tópico	Fecha	Actividades
1	Introducción y motivación al curso. Repaso de conceptos: probabilidad, probabilidad condicional.	09-13 enero	
2	Variables aleatorias. Covarianza, correlación, entropía. Distribuciones: estadísticos y resúmenes.	16-20 enero	
3	Métodos exploratorios. Visualización y dependencia entre variables. SVD. PCA.	23-27 enero	
4	Interpretación de PCA. Ejemplos y aplicaciones. Variantes de PCA. Re-escalamiento multidimensional.	30 enero-03 febrero	
5	ICA. Factoración de matrices no-negativas. Funciones kernel. Distribuciones empíricas.	06-10 febrero	
6	Métodos locales: Isomap, t-SNE, *Local Embedding*, *Manifold Learning*, SOM. *Spectral Embedding*.	13-17 febrero	
7	Métodos de agrupamiento jerárquico. Dendrogramas. $K-$medias, $K-$medianas, $K-$medoides.	20-24 febrero	
8	Mezclas gaussianas. Agrupamiento espectral: vector de Fiedler, NCuts.	27 febrero-03 marzo	
9	Métodos basados en densidades: Mean-shift. Métricas para métodos de agrupamiento.	06-10 marzo	
10	El método de $K-$vecinos más cercanos. Clasificador *Naive Bayes*.	13-17 marzo	
11	Análisis discriminante (LDA). Clasificadores lineales: el clasificador logístico.	20-24 marzo	
12	Buenas prácticas de visualización de datos.	27-31 marzo	
	Semana Santa	03-07 abril	
13	Seminario de proyectos aplicados.	10-14 abril	Proyecto 1
14	Árboles de decisión. Entropía e impureza. *Random forests*. Modelos ensamblados.	17-21 abril	
15	Modelos ensamblados: *Bagging*, *Boosting*, *Stacking*. El modelo de regresión lineal ordinaria (OLS).	24-28 abril	
16	Gráficos de diagnóstico. Selección de modelos: AIC, BIC. Mètricas de evaluación.	01-05 mayo	
17	Introducción a las redes neuronales artificiales.	08-12 mayo	
18	Seminario de proyectos aplicados.	15-19 mayo	Proyecto 2