#EspecializaciónUBA #CienciaDeDatos #Balance

[Parte 1]

Se acerca el fin de año y es tiempo de balances, en mi caso particular terminando el año pasado decidí meterme a aprovechar mi juventud y estudiar un poquito más.

La especialización en ciencia de datos de la UBA es el primer año de la Maestría y no se llama ciencia de datos, Explotación de Datos y Descubrimiento del conocimiento (<http://datamining.dc.uba.ar/datamining/>) un nombre tradicional que da cuenta de una maestría que lleva muchos años activa, muchos más que los que lleva tan de moda el nombre para todes en el último tiempo. Decidí escribir esto a fin de año por un lado para orientar a quienes estén buscando seguir estudiando acerca de que van a encontrarse. Por el otro para orientarme a mí mismo acerca de todo lo que absorbí en tan poco tiempo.

Lo primero es lo primero.

A fin del año pasado curse 3 mini materias introductorias con examen para lograr la admisión. Como ingeniero la materia de algoritmos básico y de estadística mucho no sumaban a mi formación (Aunque el repasito de estadística es FUNDAMENTAL después de tantos años) pero la mini materia de bases de datos en las que rendimos por un lado modelos relacionales y por el otro SQL fue todo nuevo. Una muy buena introducción yendo en 3 semanas bastante más lejos que cualquier curso que encontré online pero sabiendo que hay una materia en segundo año que va a fondo.

Si no sos del palo de la ingeniería tenes que saber que la mayoría de mis compañerxs tampoco, me encontré con gente de ciencias políticas, ambientales, abogacía, económicas, ingeniería, Biología, Física y cualquier rincón de la UBA. El curso introductorio nos dio la base para arrancar el primer año con conocimientos de base para poder mezclar con lo que ya traemos del grado a la mesa.

Primer Semestre (A entender las bases)

El primer semestre tiene 3 materias: Data Mining, Aprendizaje automático y Análisis inteligente de Datos.

Con estas tres materias se cubre todo lo que encontré en internet como básico para meterse en este mundo.

Data Mining (DM) se dio con un formato muy interesante en el cual íbamos compitiendo contra el profesor en Kaggle, de esta forma a medida que iban pasando las clases teníamos más herramientas para poder hacer limpieza y pre procesamiento de los datos manteniendo siempre un modelo no optimizado pero que usábamos todxs el mismo. La materia nos da un problema de clasificación grupal y uno de regresión individual y los contenidos son muy claros.

Lenguaje: Python en Colab.

Aprendizaje Automático es la otra cara de la moneda de DM, aquí nos daban una dataset ya limpio y en buena parte pre procesado y la idea era entender los diferentes modelos y el sustento teórico o la aplicación práctica. Teniendo la posibilidad de programar un árbol y ver diferentes modelos de las librerías de SciKit, hacer optimizaciones de modelos y chocarnos de frente con los problemas del overfitting.

Lenguaje: Python en Colab.

#EspecializaciónUBA #CienciaDeDatos #Balance

[Parte 2]

En esta parte vamos a hablar de las dos materias mas científicas de la maestria, AID en el primer semestre y Data Mining en Ciencia y Técnica en el Segundo

Análisis Inteligente de Datos (AID)

AID es la materia distinta, tiene un enfoque mucho mas matemático, con un libro asociado. Se da en R, tiene una introducción a la visualización, el clustering, arboles jerárquicos, diferentes diagramas (Box-plot, histograma, etc) análisis de correspondencia, test de hipótesis aplicado y muchas otras herramientas. Es una materia más descriptiva y con un enfoque más tradicional, pequeños ejemplos y una evaluación en base a TPs para poder utilizar lo aprendido por nuestra cuenta.

El libro de la materia es MUY completo, con código en r para todo y un material de referencia para cualquier persona que se quiera meter a entender a fondo información.

El libros es Análisis inteligente de datos con R : con aplicaciones a imágenes / Débora Chan, Cristina Inés Badano y Andrea Alejandra Rey. - 1a ed . - Ciudad Autónoma de Buenos Aires : edUTecNe, 2019. Es de la UTN y muy recomendable lo pueden descargar entero de la repo: <https://ria.utn.edu.ar/xmlui/handle/20.500.12272/4371>

Lenguaje: R en RStudio.

Data Mining en Ciencia y Técnica

Por ultimo DMCyT es un enfoque muy científico de la ciencia de datos, el plantel docente se dedica a la investigación y el temario es mayoritariamente aprendizaje no supervisado. Así como en Economía se utiliza un dataset tabular y un target en esta materia el enfoque está en los métodos de clustering (K-means, k-medoides, PAM, c-means, etc) y una parte muy interesante de análisis de redes (no confundir con el método de redes neuronales). El análisis de nodos y grafos para entender desde redes sociales a interconexiones del cerebro es muy interesante y es un tema que solo se ve en esta materia.

Se evalúa con dos tps, nos tocó entender los diferentes métodos de clustering analizando imágenes de flores y diferencias de forma significativa estadíos del sueño con el estudio de redes en un cerebro.

Ambos Tps fueron en colab y se escriben en Latex con lo cual en esta materia se añade la herramienta de Overleaf para escribir en grupo.

#EspecializaciónUBA #CienciaDeDatos #Balance

[Parte 3]

Si no viste las primeras partes, esto es un balance de la especialización en ciencia de datos de la UBA. Tiene 3 niveladoras y 6 materias para cursar durante el año. El trabajo de especialización quedara para el balance del año que viene. Para saber sobre las niveladoras y el primer semestre pueden ver la parte 1.

Ya sabemos lo básico, ¿que nos queda? Todo.

En el segundo semestre todo subió fuerte en intensidad, veníamos de un semestre agitado para ser un posgrado, pero manejable. Ahora ya con las herramientas en mano entramos en materias que iban a ir más profundo al problema de la ciencia de datos. Ya en la primera semana vimos que iban a haber dos materias contradictorias y complementarias Data Mining en Economía y Finanzas y Enfoque estadístico del Aprendizaje. Por otro lado, Data Mining en Ciencia y Técnica iba a ir a su propio temario por otro camino.

¿Qué problema se nos plantea desde el día 1? ¿Quiero explicar o quiero predecir?

Repasemos todas las materias primero y luego como iban chocando.

Data Mining en Economía y Finanzas

No se dejen engañar por el nombre, esta materia tiene un contexto financiero en el cual el problema que se plantea es predecir las bajas de un banco con 2 meses de anticipación. Economía y financias termina en el problema y la materia empieza en cómo resolverlo.

Es probablemente la materia mas completa en pedagogía que me toco cursar. Tiene un foro propio (Plataforma Zulip) el cual está muy incentivado a ser usado y el cual el plantel docente está muy atento. Plantea un entorno de trabajo símil a un banco, vos sos un científico de datos nuevo y tenes los scripts del banco anterior a que llegues y no vas a inventar la rueda, vas a hacer minería de datos. Con las diferentes herramientas que tenes aprendes a usar GIT, Google Cloud Platform, y Data.table dentro de R. A medida que avanza la materia entender que como evaluar un modelo, como definir si va a ser útil a futuro aun no teniendo información y por sobre todas las cosas entendes como hacer experimentos y evaluarlos para obtener información nueva.

La materia te va llevando desde lo mas básico, un solo mes de entrenamiento y RCART, hasta modelos stackeados con lightbm y Feature Engineering que cubre meses y meses de historia, arregla la base de datos, los cambios mes a mes y es optimizado paso por paso.

Incluye 4 competencias de Kaggle, una competencia de programación para entender la comparación de modelos, un video a una gerenta general y un video a un gerente técnico y una presentación individual de un experimento.

Son 3 materias, el semestre vino intenso.

Lenguaje y herramientas: R en Rstudio, Git y github, Zulip, Google Cloud Platform (Usando los 300 dolares gratis o 400 si tenes cuenta universitaria).

#EspecializaciónUBA #CienciaDeDatos #Balance

[Parte 4]

Enfoque Estadístico del Aprendizaje:

Otra materia con mucha dedicación de parte del titular. Toda la materia tiene teóricas grabadas y clases presenciales en las cuales se discute la teorica. En esto es muy bueno el profesor que te lleva a terminar de entender las limitaciones de la ciencia de datos y la fina línea entre la estadística y la ciencia de datos.

La idea de la materia es una contra cara de la de Economia, aca predecir es menos importante. Lo importante es poder explicar la información. Los modelos en general son mucho mas simples pero con una mayor capacidad explicativa. Se puede buscar una significancia, un p-valor, etc.

La materia se centra en el modelo lineal y lo va complejizando hasta llegar a las redes neuronales. El transfondo matemático es muy interesante.

Se evalúa con un TP, un choice teorico y una clase sobre un tema que complemente a la materia. La clase se graba y si queres la pueden subir al youtube. Con Flor hicimos una clase sobre Cuadrados Mínimos Recursivos que pueden ver en:

<https://www.youtube.com/watch?v=4zfRVwgmgvw> (Intro)

<https://www.youtube.com/watch?v=w9VWOZ_xGmU> (Teórica)

<https://www.youtube.com/watch?v=-GUMwuv482k> (Practica)

En el mismo canal pueden ver el de varixs compañerxs así como todas las teóricas de la materia.

Lenguaje: R en Rstudio.

¿Porque son dos caras? Si llegas a cursarlas vas a darte cuenta, la forma de explicar y entender a la información es filosóficamente distinta, al nivel de diferencia entre la ciencia física y la ingeniería. Dicen lo mismo, nadie niega de sus limitaciones ni le da poderes mágicos a lo que enseña, pero está claro que son distintas.

Entender ambas materias abre la mente.

¿Qué me llevo de este año?

Si bien me recibí en 2021 la realidad es que no curso desde antes de la pandemia. Ya en 2019 curse poco por lo que mi último año de cursada real fue 2018. Extrañaba el desafío mental de entender nuevos conceptos y este año me dio el estímulo que buscaba.

Creo que entender ciencia de datos y/o estadística es fundamental para cualquier área del saber, es un complemento para todas las carreras de grado y una forma de adelantarse a una tecnología que se está escuchando cada día más fuerte. Los modelos predictivos y explicativos son un estándar que se está asentando en todas las industrias y que tienen mucho camino por recorrer. Me alegra tener una visión mucho más clara de que es lo que hay detrás, como se evalúan y como se ponen en producción.

Por lo que me dijeron el segundo año, si decido hacerlo, es mucho más tranquilo. Enfocado en escribir los trabajos de especialización y maestría con 2 materias electivas en cada semestre es más liviano que este año.

Este fue mi balance de este año de estudio, si alguien quiere info, mis DM están abiertos.

Saludos!