Agárrense de su sillón, porque ya se viene la lista de libros que cualquier Data Scientist debería tener en su radar, por cuál empezar y que secuencia tomar, ésta lista es la que considero troncal (pudiera haber más libros de ramas específicas, por ejemplo Bayesian Analysis, Time Series, etc.. pero esas ramas las dejaremos para otro post), aunque después de leer estos libros, sin duda tendrán un conocimiento general de ramas específicas y les será sencillo adentrarse a temas particulares.

La base siempre son las matemáticas.

Un vistazo general

Mathematics for Machine Learning (Deisenroth, Faisal, Ong)

A pesar de que hay que entender que el machine learning es una herramienta dentro de la ciencia de datos, este libro sienta las bases de temas como álgebra lineal, geometría analítica, calculo vectorial, probabilidad y optimización.

Una vez conociendo las bases y estando empapados del las bases matemáticas en lo general, vamos a adentrarnos a las diferentes ramas matemáticas en las cuales deberemos desarrollar un conocimiento profundo para poder encarar temas complejos:

Las bases con más detenimiento una por una:

Linear Algebra and Learning from Data (Strang)

Introduction to Probability (Bertsekas and Tsitsiklis)

All of Statistics (Wasserman)

All of Nonparametric Statistics (Wasserman)

Estos cuatro libros no son un juego, para entrarles hay que venir con cálculo integral y diferencial a tiro, el libro de Strang nos servirá para entender de forma específica el lenguaje con el que manipulamos datos en un computador, las matrices.

Por otro lado, el libro de Bertsekas y Tsitsiklis desarrollará el lenguaje necesario para poder adentrarnos propiamente a la estadística, básicamente Strang y Tsitsiklis son requisito para poder pasar a los libros de Wasserman.

Uno nos ayudará a entender el lenguaje del que parte la estadística que es la probabilidad, (OJO no es lo mismo probabilidad y estadística, y personalmente muchos errores que veo en planteamientos o en las asesorías en DataLab es que no se tiene muy claro su diferencia y de ahí se cimientan muchos errores) y cómo hacemos para traducir un enunciado probabilista o formular un estadístico a un lenguaje que la computadora pueda consumir, que es el álgebra lineal.

Los libros de Wasserman son bastante pesados, sin embargo cuando uno le pierde el miedo al lenguaje matemático formal (que en este punto ya desarrollamos con los libros 1 y 2 de esta sección), son libros autocontenidos y súper concisos, van directo al grano.

Afinando la puntería: Saber hacer experimentos

Una vez que tenemos las bases dominadas, tenemos que entender lo siguiente, cualquier set de datos se puede entender de dos formas:

Como el resultado de un experimento consumado

Como un espacio experimental

Nos centraremos por ahora en el primer punto, cuando se entiende a los datos como el resultado de un experimento consumado, hay que entender el experimento para poder empezar a hacer preguntas inteligentes y poder sacar conclusiones de esos datos, además entender bien el proceso generador de los datos abrirá la puerta a usarlos como un espacio experimental, que no es otra cosa que a plantear nuevos experimentos dentro (con el mismo set de datos) y fuera (experimentos complementarios que nos ayudaran a entender mejor el fenómeno en cuestión).

Como resultado de un experimento consumado:

Statistics for Experimenters (P. Box)

Introduction to Econometrics with R (Hanck, Arnold, Schemelzer)

Causal Inference in Statistics A Primer (Pearl)

Causal inference for Statistics (Rubin)

Data-Driven Science and Engineering (Brunton)

Data Science for Business (Provost)

Estos libros nos ayudarán a entender como hacer set-up de experimentos y sobretodo a tratar de explicar como se generaron los datos que estamos observando, la mayoría de las técnicas que encontrarán en estos libros, son útiles en el contexto de cuando tienes un pregunta a priori sobre los datos, es decir quieres interrogar al fenómeno y asumes que la respuesta puede encontrarse ahí.

Ojo, aquí tomaremos al libro de Schemelzer no como una introducción a R si no que podremos empezar a interactuar con las preguntas que nos irán surgiendo en un espacio computacional, además de que presenta la teoría de muchos temas de una forma súper organizada y digerible.

Además el libro de Provost nos ayudará a entender las limitaciones de algunos modelos en el contexto del negocio y cosas que hay que tener en cuenta al momento de comunicar resultados, empezar un proyecto, etc.

Como un espacio experimental:

Llega la hora de empezar a plantear nuevos experimentos y ya tenemos las bases de como hacerlo!

Debemos tratar de empezar a hacer inferencias de cosas no observadas de forma directa o con una hipótesis a priori, es decir, aquí sacamos nuestro gorro de detectives y nos adentramos al mundo del aprendizaje estadístico.

An introduction to Statistical Learning (Hastie)

The Elements of Statistical Learning (Hastie)

Computer Age of Statistical Inference (Hastie)

Estos libros no son de estadística (para ver la diferencia abran los de Wasserman a un lado de estos y vean los temas) utilizan técnicas estadísticas en sets de datos complejos, algunos modelos del aprendizaje estadístico han sido absorbidos por el ML (Machine Learning) en la práctica, sin embargo existen algunas diferencias sutiles que podemos abordar en otro post.

Básicamente aquí podremos entender la relación estrecha que tienen el ML y la estadística.

Lo que todos esperaban: Machine Learning

Llegamos al punto dónde queremos automatizar el proceso de aprendizaje sobre un set de datos, cuando el problema escala y es tan complejo que establecer set de reglas o hipótesis es una tarea que suena muy muy complicada, llega el ML para salvarnos.

Pattern Recognition and Machine Learning (Bishop)

Machine Learning: A Probabilistic Perspective (Murphy)

Estos dos libros son excelentes en plantear mucho del campo de Machine Learning sin dejar por un lado su link con las partes teóricas, no son libros de implementaciones aun no llegamos a hacer código... seguimos aprendiendo el: qué, cómo, cuándo y dónde de la ciencia de datos.

¿Ya se desesperaron? aquí empieza la parte en donde usan código :) !

Algo que es de notar, es que el viaje hasta aquí es laaaaargo, y aun no hemos tocado libros con implementaciones particulares, ¿por qué?

La ciencia de datos tiene una dirección que todo DataLaber <3 debe tener clara:

Contexto -> Matemáticas -> Herramientas

Hemos estado abordando las primeras dos, la última, es la programación la cuál en el contexto de ciencia de datos es una herramienta, para poner a trabajar todas las mates que sabemos en relación a una hipótesis en un contexto particular.

Los dos primeros libros aquí sin lugar a dudas serán:

Python Data Science Handbook (Vanderplas)

R for Data Science (Wickham)

Python Y R serán sus mejores aliados en este camino (en DataLab no apoyamos esas dicusiones sobre cuál es mejor, cada uno tiene sus ventajas y desventajas, pero se complementan y no saberlos en estos días como científico de datos, es prácticamente tirar dinero a la basura), así que vale más aprender ambos, estos dos libros nos ayudarán a entender estos lenguajes de programación y a empezar a manejar datos con ellos.

¿¡Dónde está mi Machine Learning!?

Tranquilos! las implementaciones de los modelos que en este punto ya van a entender perfectamente o por lo menos, tendrán las herramientas para entenderlos se encuentran muy bien descritas en estos dos libros:

Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow (Géron)

Hands-on Machine Learning with R (Boehmke)

Sin embargo, posiblemente en la práctica van a querer mejorar más sus modelos o poderlos explicar mejor en términos de sus variables, eso se encuentra en:

Feature Engineering and Selection (Kuhn)

Interpretable Machine Learning (Molnar)

Cuando la gente se aburre del ML clásico lo que quiere son neuronas o comportamientos

Para entender el subcampo del ML dedicado al aprendizaje profundo o el aprendizaje por refuerzo los libro no son otros que:

Deep Learning (Bengio)

Reinforcement Learning (Barto)

Estos libros les ayudarán a entender la parte teórica de estos subcampos del ML aunque rara vez son técnicas recomendadas para usar en un problema de negocios en primera instancia, son parte fundamental del estudio de la inteligencia artificial y en casos específicos son las técnicas que mejor resolverán los problemas en comparación al ML clásico.

Avisos parroquiales:

Si te sirvió el contenido de este post, compartelo con tus amigos, invita a tus amigos al grupo de DataLab para que formen parte de la comunidad!

Cómo pueden ver, la ciencia de datos es un camino largo, y aun en esta lista hay muchos libros que faltan de temas más específicos, pero si leen estos de inicio a fin por lo menos una vez y los mantienen como libros de referencia, tendrán en su repertorio de científico de datos lo esencial, si sobreviven a este camino (solo falta querer - pero de verdad querer - hacerlo), cualquier tema en específico les será fácil de asimilar y de traducir al contexto en el que lo quieran aplicar.