Métodos Numéricos
Primer Cuatrimestre 2016
Trabajo Práctico 1

Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Hoy te convertís en (Junior) Data Scientist

Contexto y motivación

Comparadas con otros medios de transporte tanto de carga como de pasajeros, las aerolíneas poseen una historia relativamente corta. Con poco más de 100 años desde la creación de las primeras aerolíneas[1], los avances tecnológicos a lo largo del Siglo XX permitieron el continuo desarrollo de la industria aeronáutica. La importancia de la misma radica no solo en la posiblidad de acortar distancias y tiempos de viaje, sino que también tiene un impacto muy importante a nivel social, científico y económico. Como contraparte, la operatoria diaria presenta un nivel de dificultad muy alto. Esto incluye grandes estructuras organizativas y de infraestructura que demanda inversiones considerables a mediano y largo plazo, altos costos operativos y la garantía del cumplimiento de medidas de seguridad muy complejas, entre otras cosas.

En términos generales, las grandes organizaciones (empresas privadas, organizaciones públicas, gobiernos, etc.), necesitan realizar evaluaciones periódicas respecto de sus actividades con el fin de establecer si se encuentran funcionando correctamente, determinar si se han alcanzado las metas generales propuestas e identificar posibles puntos de conflicto. Una posible forma de llevar a delante esta práctica es mediante los llamados *indicadores de performance* (KPIs, por su nombre en inglés, *Key Performance Indicators*) que consisten en métricas asociadas a actividades particulares dentro de la organización. Existen distintos tipos de KPIs (cualitativos, cuantitativos, etc.) y que apuntan a distintos aspectos de una organización (financiero, operativos, de manufactura, de gobierno, etc.).[2]

El problema

Como se menciona anteriormente, las aerolíneas son organizaciones naturalmente complejas en muchos niveles distintos, ya sea a nivel financiero y la sustentabilidad de la compañia como a nivel de satisfacción del cliente y su percepción del servicio brindado. Existen diversos KPIs que pueden que evaluar distintos aspectos: financieros, operativos, organizacionales, etc. A modo de ejemplo, los indicadores principales considerados por British Airways y como son medidos pueden verse en [1, 2]. A nivel operacional, en [3] se explica detalladamente algunas de las métricas, junto con su problemática e importancia, utilizadas en Estados Unidos y Europa, y se realiza una comparación entre ellas.

Uno de los KPIs utilizados para evaluar las operaciones en sistemas de transporte, en particular para aerolíneas, es la puntualidad (OTP, conocidos también como *punctuality*, u *On-Time Performance*) de los servicios. Para la industria aeronáutica en particular, muchas de las

[1] KLM y Qantas volaron por primera vez en 1920, mientras que Aerolńeas Argentinas fue creada en 1950.

[2] Es importante destacar que la definición de los mismos, las métricas a utilizar y las acciones a tomar en función de los resultados dependen de los criterios de la organización, y pueden ser utilizados de forma incorrecta.

decisiones a tomar en la planificación de las operaciones diarias se realizan en base a las programaciones de horarios de las aerolíneas. En este sentido, es importante destacar que estas decisiones son interdependientes y que involucran no solo a las aerolíneas, sino también a al aeropuerto y sus prestadores de servicios que debe planificar la utilización de sus (acotados) recursos.

Actualmente, un vuelo se considera retrasado (*delayed*) si su arribo (o partida) se produce 15 minutos después de lo planificado en la programación original. Las operaciones diarias son influenciadas por eventos inesperados que no siempre son posibles de evitar y por lo tanto es habitual tener vuelos con delay. Medir efectivamente la magnitud de los mismos permite realizar una evaluación sobre como han sido las operaciones hasta el momento. Más aún, tener una herramienta de predicción confiable para establecer la magnitud del fenómeno en el futuro, que puede ser utilizada mediante técnicas de optimización para tomar decisiones y eventualmente realizar modificaciones en las programaciones.

El indicador OTP es particularmente interesante ya que afecta directa e indirectamente distintos aspectos. Por ejemplo, la presencia de gran cantidad de delays significativos afectan la utilización de los recursos del aeropuerto en función de su planificación original, pudiendo generar cuellos de botella en la misma y afectar indirectamente las puntualidad de otros servicios. Este factor es crítico en escenarios con una demanda intesiva de recursos de capacidad limitada. Más aún, esto puede traducirse en un incremento considerable de los costos operativos, por excederse en uso de recuros (pista, manga, etc.) y por penalizaciones. Por otro lado, afecta directamente a la percepción de los usuarios respecto a la calidad del servicio brindado, ya que los retrasos pueden provocar no solo tiempos de espera más largos, si no también, la pérdida de vuelos en conexión.

El trabajo práctico consiste en aplicar técnicas de Métodos Númericos y *Data Science*, en particular Regresiones Lineales/Cuadrados Mínimos. Estas técnicas serán aplicadas a un (gran) conjunto de datos reales buscando proveer información descriptiva y modelos que puedan ser utilizados para predecir fenómenos relacionados, pero no necesariamente limitados a, OTP. Los datos comprenden cieta información relacionada a vuelos realizados en Estados Unidos entre los años 1987 y 2008, incluyendo información de la compañia, fecha y horarios planificados de partida/arribo, horarios reales de salida/llegada, causa del delay, si fueron cancelados o no, y su respectiva causa, el tipo de avión utilizado, tiempo de vuelo, tiempos de *taxi*, entre otras cosas. Los mismos deben ser obtenidos de la competencia organizada [4], donde además se incluye una descripción de cada campo. La misma se centró principalmente en visualización y análisis de datos, aunque no tanto en predicción.

Uno de los principales objetivos del trabajo práctico es que cada grupo pueda aplicar parte del conocimiento metodológico adquirido durante los primeros dos tercios del cuatrimestre. Por este motivo, los grupos deberán proponer aspectos a analizar de los datos y formular los experimentos necesarios, siguiendo los lineamientos y requerimientos mínimos pre-establecidos.[3] Esto se debe además a que se busca que los grupos puedan realizar distintos estudios entre sí, identificando diferentes líneas de experimentación y análisis.

Con el fin de guiar los posibles ejes de estudio, se presenta a continuación algunos posibles disparadores:

[3]Se recomienda a los grupos validar con los docentes las ideas y direcciones elegidas en las clases de consulta.

- Cómo varía la cantidad de vuelos cancelados por mes a través de los años? Y la magnitud de los retrasos?
- Es posible caracterizar la cantidad de vuelos cancelados y/o magnitud de los delays en función de dia/mes? Que nivel de granularidad en función del tiempo es conveniente tomar?
- Todos los aeropuertos se comportan de la misma manera? Y las compañias aéreas? Y entre pares de ciudades en particular?
- Es importante diferenciar efectos estacionales (clima, temporada alta, fechas particulares con picos de demanda, etc.)?
- El tipo/antigüedad en los aviones es importante?
- Las condiciones y requerimientos mínimos de seguridad produjeron cambios significativos luego del 9/11. Como afecta esto a los modelos predictivos?

Técnicas a utilizar y métricas de evaluación

La técnica de Métodos Numéricos a utilizar para proponer los modelos es Regresiones Lineales/Cuadrados Mínimos Lineales (CML). Para determinar nuestro modelo, asumimos tener una serie de N observaciones $(x_{(i)}, y_{(i)})$, con $x_{(i)} \in \mathbb{R}^k$ el vector de *features* e $y_{(i)} \in \mathbb{R}$ nuestra variable dependiente. Luego, el modelo consiste en encontrar los parámetros (lineales) que definen $y_{(i)} = f(x_{(i)}) + \epsilon_i$, $i = 1, \ldots, N$, donde ϵ_i es el error de la medición i-ésima, y que minimizan el error de la aproximación en el sentido de CML.

Dado un conjunto de datos $\{(x_{(i)}, y_{(i)}\}_{i=1,\ldots,N}$ será necesario considerar distintas hipótesis sobre la función f (por ejemplo, considerar polinomios de distinto grado) que dan lugar a distintos modelos. Para poder decidir entre los mismos, tenemos que considerar alguna métrica de evaluación. Se sugiere como mínimo considerar el *Mean Squared Error* (MSE)[4]. Dado un modelo $\hat{f}$ de f y una observación $(x_{(i)}, y_{(i)})$, definimos $\hat{y}_{(i)} = \hat{f}(x_{(i)})$ y $e_{(i)} = y_{(i)} - \hat{y}_{(i)}$. Con estas definiciones, podemos calcular el MSE del modelo $\hat{f}$ como

$$MSE(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} e_{(i)}^2.$$

Esta metodología nos sirve para evaluar cuan bien ajusta el modelo en función de los datos de entrenamiento utilizados. Sin embargo, de forma similar que en el trabajo práctico anterior, en un contexto de modelos predictivos se corre el riesgo de caer en el conocido *overfitting*. Para evitar este fenómenot, nuevamente, podemos considerar la técnica de *cross-validation* (CV). Para ello, podemos particionar nuestro conjunto de datos y variar la composición de la base de entrenamiento (*training*) y las observaciones consideradas como *test*[5]. Una vez obtenido el modelo $\hat{f}$, se toman las observaciones en el conjunto de test, se aplica el modelo y

[4] Notar que MSE es dependiente de la escala.

[5] Nuevamente, en este caso, tener en cuenta cómo afecta a variables cuyo valor dependa de la cantidad de observaciones tomadas.

se evalúa el MSE obtenido. El MSE final para el modelo $\hat{f}$ consiste en tomar alguna medida sobre los resultados obtenidos para cada combinación de training/test considerado.

Por las características de los datos disponibles y los posibles ejes de análisis, muy posiblemente sea necesario asumir que las variables a estimar no son completamente independientes y que existe una relación entre ellas. Un claro ejemplo de esta situación se da con las denominadas *series de tiempo*, donde los datos presentan un ordenamiento temporal natural.En este contexto, la metodología de evaluación es similar pero el conjunto de datos de entrenamiento solo puede considerar datos que ocurrieron previamente. Para ello, consideramos que cada observación está asociada a un determinado período de tiempo t, con $t = 1, \ldots, T$, $(x^t_{(i)}, y^t_{(i)})$, y asumimos que al menos K períodos de tiempo son necesarios para poder conformar el conjunto de *training*. Para evaluar los resultados de la predicción en el período $\tau[K, T]$ se puede:

1. Tomar los conjuntos de observaciones correspondientes a períodos $1, \ldots, \tau - 1$ como training.
2. Calcular las métricas correspondientes tomando como test el período τ.
3. Al finalizar, reportar alguna medida sobre los resultados parciales obtenidos.

El procedimiento presentado puede ser modificado. Por ejemplo, si se considera que datos muy lejos en el horizonte de tiempo no son representativos es posible restringir cuantos períodos previos considerar para el training. A su vez, la evaluación respecto a la calidad de la predicción puede considerar más de un período futuro.

Enunciado

El Trabajo Práctico consiste como punto de partida considerar los datos provistos por [4] y formular distintos ejes de análisis relacionados con la temática propuesta, OTP. Para ello, se deberá utilizar CML como técnica de análisis y modelado, tanto a nivel descriptivo de los datos como a nivel predictivo de eventos futuros. Para el desarrollo de los métodos se podrá considerar como posibles lenguajes MATLAB, Python y/o C++. Se remarca que, a diferencia de trabajos anteriores, no es necesario realizar toda la implementación desde cero y es posible utilizar rutinas provistas por dichos lenguajes. El objetivo principal de este trabajo se centra en la aplicación de las técnicas de CML a una temática práctica concreta y en la correspondiente experimentación necesaria para evaluar los desarrollos. Se deben reportar dos ejes de análisis de forma completa, donde al menos uno tiene que ser original del grupo y no puede ser exactamente alguno de los disparadores provistos por la cátedra.

El set de datos total de datos contiene más de 120 millones de registros, divididos en un conjnto de archivos en función del año de los mismos, ocuapdo aproximadamente 1.6 Gb comprimidos. Por esta razón, es importante contar con herramientas sencillas que perminta extraer la información de interés para el grupo. Junto con este enunciado se entregan algunos ejemplos que utilizan comandos báscios de scripting (`awk, cut, grep, wc`) para realizar operaciones útiles de filtrado de datos. Desde ya que su utilización no es obligatoria, y se inivta a los grupos a extenderlos o incluso utilizar otras herramientas. En este sentido, es posible que los grupos compartan, a través de la lista de alumnos de la materia, herramientas de prepeocesamiento y extracción de datos con otros grupos. Es importante evitar que las herramientas compartidas contengan información particular de los ejes de análisis y la experimentación a realizar.

Los resultados deben ser volcados en un informe con la estructura habitual. Sin embargo, en este caso es obligatorio escribirlo utilizando el template de la revista *Electronic Notes on Discrete Mathematics* (ENDM). Además, el informe no podrá exceder las 10 páginas de longitud, y por lo tanto los resultados tienen que ser presentados y condensados de forma adecuada. Notar que esto no significa que la experimentación debe ser acotada, si no todo lo contrario: es importante realizar muchos experimentos y mostrar los que resulten representativos. Como en los demás trabajos, es importante proveer la información necesaria para poder replicar todos los experimentos, ya sean que se encuentren en el informe o no.

Por último, este trabajo tendrá una presentación oral frente a un grupo acotado de docentes que será evaluada como una parte adicional de la nota. Para la misma, cada grupo diseñara una presentación incluyendo los desarrollos y resultados que considere intersantes, plasmados en el informe, y dispondrá de 15 minutos para exponerlo. La exposición puede ser de la totalidad o de un subconjunto de los integrantes, y esta decisión queda a elección del grupo. Una vez finalizada la misma, se llevará a cabo un coloquio donde los integrantes del grupo responderán a las preguntas realizadas. Cabe mencionar que los docentes podrán elegir que alumno debe responder, con lo cual es importante que todos los integrantes estén al tanto de todas las decisiones tomadas.

Fechas de entrega

- *Formato Electrónico:* Domingo 19/6, hasta las 23:59 hs, enviando el trabajo (informe + código) a la dirección `metnum.lab@gmail.com`. El subject del email debe comenzar con el texto `[TP3]` seguido de la lista de apellidos de los integrantes del grupo.
- *Confirmación presentación oral:* Viernes 24/6, por correo electronico.
- *Presentación oral:* Lunes 27/6, en horario a determinar luego de la confirmación. Será en horario de clase de la materia.

Importante: El horario es estricto. Los correos recibidos después de la hora indicada serán considerados re-entrega.

Referencias

[1] British Airways. 2008/2009 annual report and accounts. (link), 2009.

[2] British Airways. 2009/2010 annual report and accounts. (link), 2010.

[3] EUROCONTROL/FAA. Comparison of air traffic management-related operational performance: Us/europe. (link), 2013.

[4] ASA Section on Statistical Computing. 2009 data expo competition. (link), 2009.