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

Cuadrados Mínimos Lineales

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 sólo en la posibilidad de acortar distancias y tiempos de viaje, sino en que también tiene un impacto muy importante a nivel social, científico y económico.

En términos generales, las grandes organizaciones (empresas privadas, organizaciones públicas, gobiernos, aerolíneas, 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 adelante 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.) que apuntan a distintos aspectos de una organización (financieros, operativos, de manufactura, de gobierno, etc.)[2].

El problema

Las aerolíneas son organizaciones naturalmente complejas en muchos niveles distintos, ya sea a nivel financiero y la sustentabilidad de la compañía como a nivel de satisfacción del cliente y su percepción del servicio brindado. Para eso, utilizan diversos KPIs que permiten evaluar distintos aspectos financieros, operativos, organizacionales, etc. A modo de ejemplo, en [1, 2] pueden verse los indicadores principales considerados por British Airways, y cómo son medidos. A nivel operacional, en [3] se explican detalladamente algunas de las métricas, junto con su problemática e importancia, utilizadas en Estados Unidos y Europa (además 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, conocida también como *punctuality*, u *On-Time Performance (OTP)* de los servicios. Para la industria aeronáutica en particular, muchas de las 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 sólo a las aerolíneas, sino también al aeropuerto y sus prestadores de servicios que deben 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. Muchas veces, 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*. Tener una herramienta de predicción confiable para establecer la magnitud de este fenómeno en el futuro, es indispensable para mejorar el servicio a los clientes y disminuir los costos de la empresa.

[1] KLM y Qantas volaron por primera vez en 1920, mientras que Aerolíneas 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.

El indicador OTP es particularmente interesante ya que afecta directa e indirectamente distintos aspectos. Por ejemplo, la presencia de gran cantidad de delays significativos afecta 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 puntualidades de otros servicios. Este factor es crítico en escenarios con una demanda intensiva de recursos de capacidad limitada. Más aún, esto puede traducirse en un incremento considerable de los costos operativos, por excederse en uso de recursos (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 sólo tiempos de espera más largos, sino también la pérdida de vuelos en conexión.

El presente trabajo práctico consiste en aplicar técnicas de Métodos Numéricos y *Data Science*, en particular Regresiones Lineales con Cuadrados Mínimos sobre un (gran) conjunto de datos buscando proveer información descriptiva y de modelos que puedan ser utilizados para predecir fenómenos que afecten a la puntualidad (OTP), pero no necesariamente limitados a ésta.

Data sets

Los datos a analizar comprenden cierta información relacionada a vuelos realizados en Estados Unidos entre los años 1987 y 2008, incluyendo información de la compañía, 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 [5], donde además se incluye una descripción de cada campo. Notar que la mencionada competencia se centró principalmente en visualización y análisis de datos, aunque no tanto en predicción (en *"Posters & results"* de [5] encontrarán los trabajos ganadores).

El set de datos contiene más de 120 millones de registros, divididos en un conjunto de archivos en función del año de los mismos, ocupando aproximadamente 1.7 GB comprimidos. Por esta razón, es importante contar con herramientas sencillas que permitan extraer la información de interés para el grupo. Junto con este enunciado se entregan algunos ejemplos que utilizan comandos básicos de scripting (`awk, cut, grep, wc`) para realizar operaciones útiles de filtrado de datos. Desde ya que su utilización no es obligatoria, y se invita a los grupos a extenderlos o incluso a utilizar otras herramientas.

Técnicas a utilizar y métricas de evaluación

La técnica de Métodos Numéricos a utilizar para proponer los modelos es el de Regresiones Lineales con 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 i-ésima medición) y que minimizan el error de la aproximación en el sentido de CML.

Dado un conjunto de datos $\{(x_{(i)}, y_{(i)})\}_{i=1,\ ...,\ N}$ 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 *Root Mean Squared Error* (RMSE). 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 RMSE del modelo $\hat{f}$ como:

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

Esta metodología nos sirve para evaluar cuán bien ajusta el modelo en función de los datos de entrenamiento utilizados. Tener en cuenta que la calidad de predicción del modelo la deberemos evaluar en los datos de testing o validación, previamente definidos (distintos a los de entrenamiento) para lo cual es posible utilizar la misma métrica.

Notar que el RMSE es dependiente de la escala, por lo tanto cuando se quiera aplicar sobre conjuntos de datos distintos utilizar su versión normalizada:

$$NRMSE(\hat{f}) = \frac{RMSE(\hat{f})}{y_{\text{máx}} - y_{\text{mín}}}$$

donde $y_{\text{máx}}$, $y_{\text{mín}}$ corresponden al máximo y mínimo de los valores del conjunto analizado, respectivamente.

También es posible considerar otras formas de normalizar o métricas de evaluación en la predicción (ver [4]).

Series de tiempo

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 sólo puede considerar datos que ocurrieron previamente. Para ello, consideramos que cada observación está asociada a un determinado período de tiempo t, con $t = 1, \ldots, T$, $(x_{(i)}^t, y_{(i)}^t)$, 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 lejanos en el horizonte de tiempo no son representativos es posible restringir cuantos períodos previos considerar para el training. A su vez, para la evaluación respecto de la calidad de la predicción se puede considerar más de un período futuro.

Experimentación

De forma similar al trabajo práctico anterior, en un contexto de modelos predictivos se corre el riesgo de caer en el conocido *overfitting*. Para evitar este fenómeno, 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*[3]. Una vez obtenido el modelo $\hat{f}$, se toman las observaciones en el conjunto de test, se aplica el modelo y se evalúa el RMSE obtenido. El RMSE final para el modelo $\hat{f}$ consiste en tomar alguna medida sobre los resultados obtenidos para cada combinación de training/test considerado.

Con el fin de guiar los posibles ejes de estudio, se presentan a continuación algunos interrogantes que servirán como disparadores para la experimentación.

- ¿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 del día/mes? ¿Qué nivel de granularidad en función del tiempo es conveniente tomar?
- ¿Todos los aeropuertos se comportan de la misma manera? ¿Y las compañías aéreas? ¿Y entre pares de ciudades en particular?
- ¿Es importante diferenciar efectos estacionales como el 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 en los Estados Unidos. ¿Cómo afecta esto a los modelos predictivos?

Enunciado

El Trabajo Práctico tiene como punto de partida considerar los datos provistos por [5] y formular distintos ejes de análisis relacionados con la temática propuesta.

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 la experimentación se podrá considerar como posibles lenguajes Python y/o C++, pero la implementación de CML **debe** ser en C++. Para la misma pueden utilizar SVD, QR o ecuaciones normales.

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 mientras no resuelvan CML. 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.

Otro objetivo 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 preestablecidos más adelante. Esto se debe a que se busca que los grupos puedan realizar y proponer distintos estudios, identificando diferentes líneas de experimentación y análisis.

Se **deben** reportar al menos dos ejes de análisis de forma completa, uno basado en el OTP (puntualidad) y otro original propuesto por los alumnos (relacionado o no con el OTP).

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

Informe

Como en trabajos prácticos anteriores, los resultados deben ser volcados en un informe siguiendo como base las pautas de laboratorio. Sin embargo, en este caso además se **deben** incluir trabajos relacionados en la introducción y citarlos utilizando la herramienta BibTeX.

También, es **obligatorio** escribirlo utilizando el template de la revista *Electronic Notes on Discrete Mathematics* (ENDM)[4]. El informe **no podrá exceder** las 10 páginas de longitud (excluyendo referencias), 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, sino todo lo contrario: es importante realizar muchos experimentos pero solamente mostrar los que resulten significativos[5]. Como en los demás trabajos prácticos, también **deben** proveer la información necesaria para poder replicar todos los experimentos (código fuente, scripts, archivos auxiliares, etc.), ya sea que se encuentren en el informe o no.

Presentación oral

Como se mencionó anteriormente, además del informe del trabajo práctico, se deberá crear una presentación con diapositivas la cual cada grupo deberá exponer en a lo sumo 15 minutos en la fecha indicada por los docentes. La presentación y exposición será evaluada y formará parte de la nota del trabajo práctico. Para la misma se deberán seguir las siguientes pautas:

- No sobrepasar los **15 minutos** de exposición.
- La exposición es requisito para aprobar el TP3 pero la misma no implica garantía de aprobación de todo el trabajo práctico.
- La exposición puede ser de la totalidad o de un subconjunto de los integrantes, y esta decisión queda a elección de cada 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 qué alumno debe responder, con lo cual es importante que todos los integrantes estén al tanto de todas las decisiones tomadas.
- Ver consejos para la creación de presentaciones en el siguiente link:
  `https://campus.exactas.uba.ar/pluginfile.php/126018/course/section/17889/ConsejosExpOral.pdf`

Fechas de entrega

- *Formato Electrónico:* domingo 23 de junio 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 separados por punto y coma `;`.
    Ejemplo: `[TP3] Lennon; McCartney; Starr; Harrison`

[4] `http://cdn.elsevier.com/promis_misc/endm_package.zip`
[5] Se podrá mostrar mayor experimentación o resultados durante la presentación oral

    - Se ruega no sobrepasar el máximo permitido de archivos adjuntos de 20MB. Tener en cuenta al realizar la entrega de no ajuntar bases de datos disponibles en la web, resultados duplicados o archivos de backup.

- *Recuperatorio:* jueves 11 de julio hasta las 23.59 hs, enviando el trabajo corregido a la dirección `metnum.lab@gmail.com`
- *Exposición oral:* viernes 12 de julio desde las 17hs.
- Pautas de laboratorio:
  `https://campus.exactas.uba.ar/pluginfile.php/126018/course/section/17889/pautas.pdf`

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] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

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