Hoy te convertís en (Junior) Data Scientist

Ricardo Colombo Diego Santos Erik Machicado

Departamento de Computación
Universidad de Buenos Aires
C.A.B.A, Argentina

Abstract

El trabajo consiste en aplicar técnicas de Métodos numéricos y Data Science, en particular Regresiones Lineales y Cuadrados Mínimos, a un gran conjunto de datos reales, buscando identificar modelos capaces de predecir comportamientos .

Keywords: Data Science, KPIs, Cuadrados Minimos Lineales (CML)

1 Introducción

En la competitiva industria aeronáutica la eficiencia de los procesos es vital para mantener la calidad del servicio. Cumplir con el servicio no es algo que solo dependa de la empresa que se contrata, ya que ao a ao la cantidad de vuelos se multipilica y muchas veces existen variables externas que terminen afectando la puntualidad y calidad del servicio. Por eso es importante contar con métricas que analizen los eventos pasados buscando patrones e intentar prevenirlos o desminuir su impacto en el futuro. Estos indicadores son denominados *Key Performance Indicators* (KPIs).

En este trabajo estudiaremos factores que influyen en la organizacin de las salidas de vuelos para un aeropuerto en particular. Como se mencionó anteriormente estadísticamente, todos los aos se registran mas vuelos para

cualquier aeropuerto, esto requiere una sincronización precisa de los tiempos que ocupa un avión dentro del aeropuerto esperando para partir.

Para la realizacin del trabajo analizaremos un set de datos reales correspondientes a vuelos realizados en Estados Unidos durante el período 1987 - 2008. Y para poder relacionar los casos utilizaremos en los dos ejes el mismo aeropuerto.

Nuestro primer eje de estudio es la evolución en la cantidad de tráfico áereo de un aeropuerto, y como la cuota de *market share* se fue concentrando con el correr del tiempo sobre las empresas líderes del segmento. Lo interesante de esta evaluación es esperamos poder proyectar el crecimiento de tráfico sobre el aeropuerto a fin de poder satisfacer la creciente demanda.

El segundo eje de estudio se centra en los factores estacionales que influyen en un vuelo salga en tiempo y forma, es decir estudiaremos la cantidad de retrasos en partidas para un aeropuerto particular, con el objetivo de detectar las estaciones anuales donde el clima pueda afectar el funcionamiento del aeropuerto.

Luego relacionaremos los ejes de estudio para estudiar si es posible predecir temporadas de mayor retraso, esperando que sea de utilidad en la programación de los nuevos vuelos que se incorporn al mercado.

Metodo de Minimos Cuadrados

Minimos cuadrados es una ténica de aná numérico, en la que, dados un conjunto de pares intenta encontrar la función que mejor aproxime a los datos de acuerdo con el criterio de Error cuadrático medio. En su forma más simple, intenta minimizar la suma de los cuadrados de las diferencias entre los puntos generados por la función de aproximación y los que corresponden a los datos. CML es el caso en el que se usa a las rectas para la aproximación osea $y = ax + b$, además la generalización del método propone encontrar la función que mejor aproxima de la forma $y = a_1 f_1(x) + \ldots + a_K f_K(x)$ y que no es necesario que las funciones f sean lineales pero si que y sea una combinación lineal de ellas. Para mas informacion consultar [**?**] (seccion 8.1)

2 Desarrollo

Como se dijo previamente, comenzamos con un set de datos relacionados a los vuelos realizados en Estados Unidos entre 1987 y 2008 (descripto en [**?**]), luego seleccionamos aquellos datos relacionados con el aeropuerto JFK, entonces propucimos como aspectos a analizar, ver el numero de vuelos con

Weather Delay mayor a 15min por mes y la cantidad de vuelos por mes de un conjunto aerolineas de similares caracteristicas (Delta, American y United). Posteriormente con los datos de esos ejes de estudio ya obtenidos para un periodo de tiempo vamos a aplicar el metodo propuesto por la catedra, dicho metodo encontrara la combinacion lineal de funciones (propuestas por el grupo) que mejor aproxime a los datos, con esta combinacion de funciones veremos si tambien aproxima a los resultados de los ejes de estudio para un periodo de tiempo posterior

Para evitar caer en el conocido overfitting elegiremos distintos subconjuntos del training para hallar los coeficientes de la aproximacion lineal.

Al momento de analizar que funciones utilizar en el metodo, como primer opcion se propuso a los polinomios, pues estos fueron estudiados y usados en clase en la introduccion del metodo. ‘riormente se plateo el uso de senos y cosenos, ya que al ser funciones periodicas podrian presentar alguna ventaja si los datos presentan patrones que se repiten varias veces.

3 Resultados

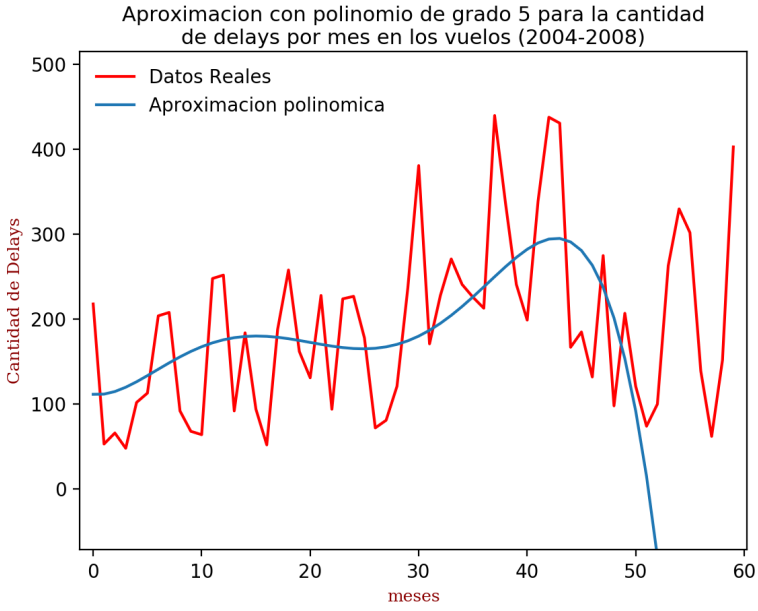
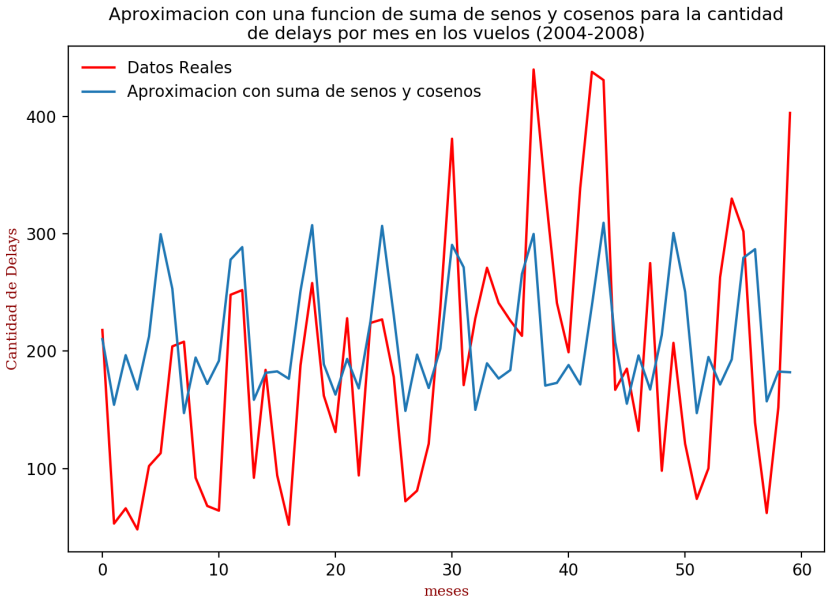
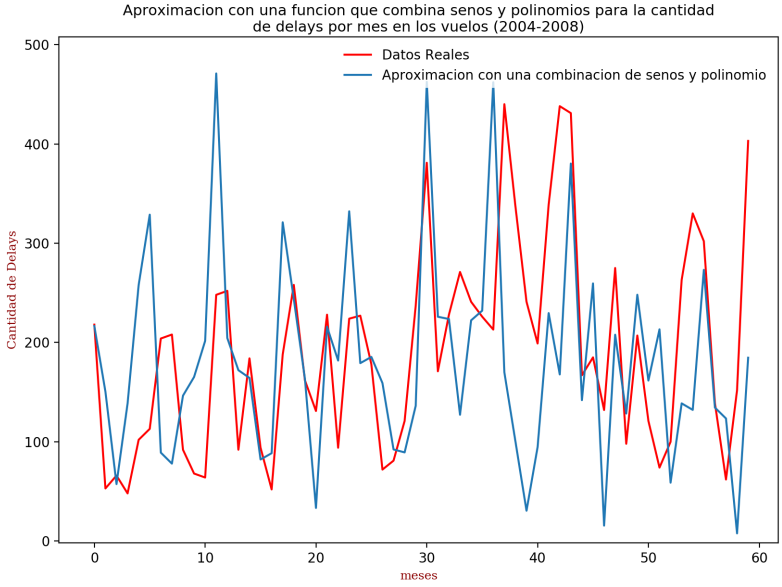
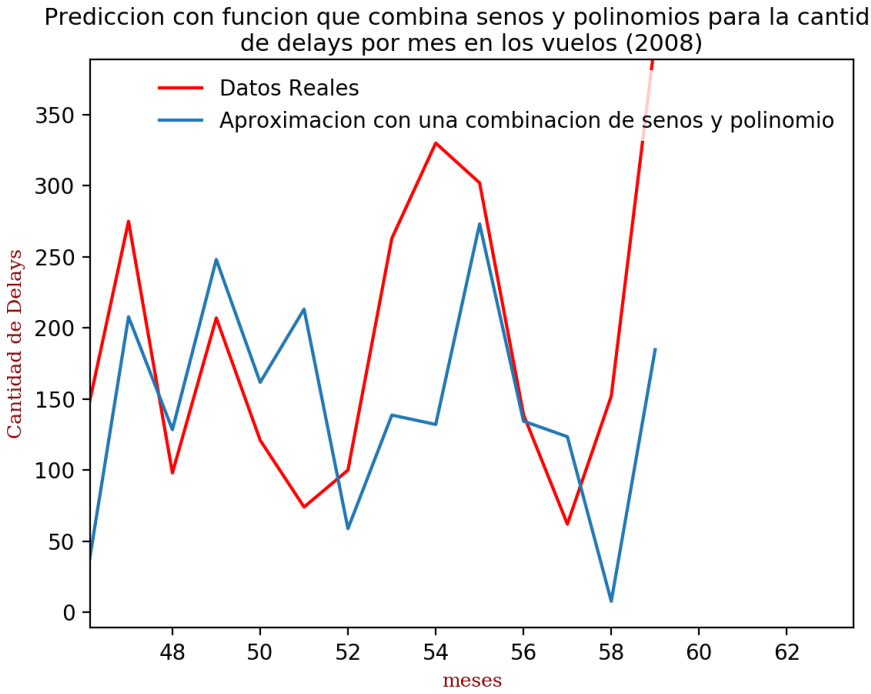


figura 1

figura 2


Aproximacion con una funcion de suma de senos y cosenos para la cantidad de delays por mes en los vuelos (2004-2008)
Datos Reales
Aproximacion con suma de senos y cosenos
Cantidad de Delays
meses
100
200
300
400
0
10
20
30
40
50
60


Aproximacion con una funcion que combina senos y polinomios para la cantidad de delays por mes en los vuelos (2004-2008)
Datos Reales
Aproximacion con una combinacion de senos y polinomio
Cantidad de Delays
meses
0
100
200
300
400
500
0
10
20
30
40
50
60



4 Discusión

Durante la experimentacion con la familias de polinomios usadas para el metodo se encontro que si bien en los meses de training se aproximaba bien afuera de ese periodo se aleja de los datos reales, como esta ilustrado en el grafico de la figura 1 que utiliza los datos del periodo 2004-2007 para predecir 2008 (apartir del mes 48) despues de discutirlo entendimos que la cantida de picos que tiene un polinomio esta limitada por el grado del mismo, pues son las raices del polinomio derivado y estas a lo sumo son tantas como el grado del polinomio derivado por eso fuera del periodo usado como training no se generan nuevos picos apesar que los datos reales si los tienen. Cuando usamos senos y cosenos como familias de funciones para el metodo nos encontramos con que estas estaban acotadas como lo estan los senos y cosenos, y no podian crecer si los datos lo hacian como lo muestra la figura 2

5 Conclusiones

References

[1] Burden, R. L., and J. D. Faires, "Análisis numérico, " Math. Surveys **7**, Amer. Math. Soc., Providence, R.I., 1961.

[2] ASA Section on Statistical Computing. 2009 data expor competition, URL: http://stat-computing.org/dataexpo/2009/the-data.html.