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.

Tomando como punto de partida los KPIs, en este trabajo se propone analizar datos reales de vuelos realizados en USA , además se busca utilizar CML para identificar modelos que describan algunos comportamientos y evaluar su eficacia para realizar predicciones. Para ello la catedra propuso que se escoja dos ejes de estudios a elección y se evalue experimentalmente

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 que 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 [2] (seccion 8.1)

2 Desarrollo

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 aerop-

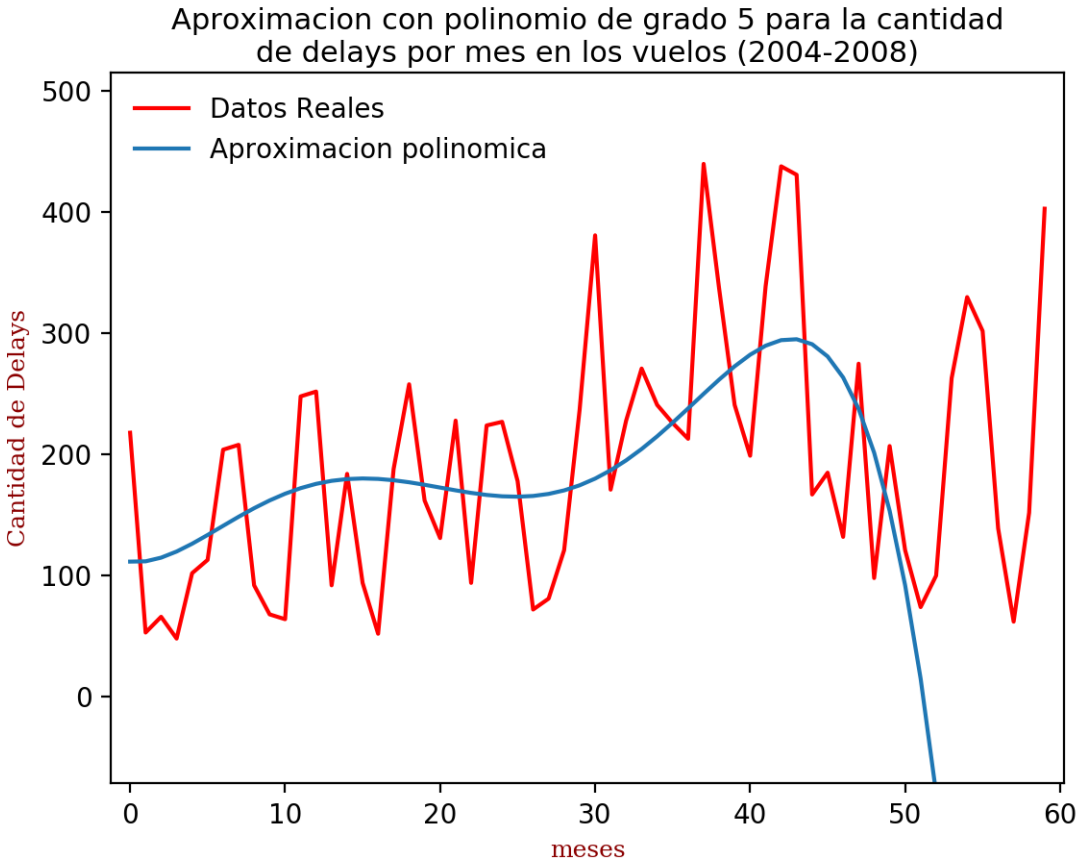
uerto. Ademas al momento de analizar que funciones utilizar en el metodo, como primer opcion se propuso a los polinomios, pues estos fueron estudiados y usados en clase en la introduccion del metodo. Posteriormente se plateo el uso de senos y cosenos, ya que al ser funciones periodicas podrian presentar alguna ventaja en la prediccion.

Cross Validation

Para evaluar los resultados de la prediccion y se siguien los siguientes pasos:

3 Experimentacion

Durante la experimentacion con la familias de polinomios usadas para el metodo se encontro que si bien en los meses tomados como training se comportaban bastante bien afuera de ese periodo crecia constantemente.



Using ENDM Macros with Mac OS X

Clearly, if your file does not require `.eps` or other PostScript files, then you can create the required `.pdf` file using any of the standard TEX implementations for the Macintosh. If you do need to include PostScript files and if you are using TEXShop, then you can specify to use dvips and Ghostview in processing your file, and then you can apply `ps2pdf` to create the needed `.pdf` file. Alternatively, the Mac OS X operating system is based on UNIX, so it supports the use of teTEX as described above.

4 Summary and Remarks

The ENDM macro package is relatively easy to use and provides a uniform layout for all the papers that appear in ENDM.

Assigning Volume Numbers

An additional point worth mentioning is that ENDM has moved to *ScienceDirect*, Elsevier's main platform for publishing electronic series. Because *ScienceDirect* cannot easily accommodate changes to published material, the *Proceedings* must be entirely ready before they can be published. Volume numbers will therefore not be assigned for the *Proceedings* until the final versions of all papers are in.

Copyright Transfer Forms

Due to the move to *ScienceDirect*, the corresponding author of each paper published in ENDM must submit a signed Copyright Transfer Form to Elsevier in order for their paper to be published. A copy of this form will be sent to each author. Note that the publication of an abstract or extended abstract in ENDM will not restrict the author(s) from publishing a full-length article on the same topic and with the same title in another journal (possibly with another publisher). Details about the copyright agreement specifying the exact rights of the authors and the rights of Elsevier are available at Elsevier's Author Gateway.

5 Bibliographical references

ENDM employs the `plain` style of bibliographic references in which references are numbered sequentially and listed in alphabetical order according to the first author's last name. Please utilize this style. We have a BibTEX style file, for those who wish to use it. It is the file `endm.bst` which is included in this package. The basic rules we have employed are the following:

- Authors' names should be listed in alphabetical order, with the first author's last name listed first followed by initials or first name, and with the other authors' names listed as *first name, last name*.
- Titles of articles in journals should be in *emphasized* font.
- Titles of books, monographs, etc. should be in quotations.
- Journal names should be in plain roman type.
- Journal volume numbers should be in boldface, immediately followed by the year of publication enclosed in parentheses in roman type.
- References to URLs on the net should be "active" and the URL itself should be in `typewriter` font.
- Articles should include page numbers.

The criteria are illustrated by the examples below.

References

[1] Civin, P., and B. Yood, *Involutions on Banach algebras*, Pacific J. Math. **9** (1959), 415–436.

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

[3] Freyd, Peter, Peter O'Hearn, John Power, Robert Tennent and Makoto Takeyama, *Bireflectivity*, Electronic Notes in Theoretical Computer Science **1** (1995), URL: http://www.elsevier.com/locate/entcs/volume1.html.

[4] Easdown, D., and W. D. Munn, *Trace functions on inverse semigroup algebras*, U. of Glasgow, Dept. of Math., preprint 93/52.

[5] Roscoe, A. W., "The Theory and Practice of Concurrency," Prentice Hall Series in Computer Science, Prentice Hall Publishers, London, New York (1198), 565pp. With associated web site http://www.comlab.ox.ac.uk/oucl/publications/books/concurrency/.

[6] Shehadah, A. A., "Embedding theorems for semigroups with involution, " Ph.D. thesis, Purdue University, Indiana, 1982.

[7] Weyl, H., "The Classical Groups," 2nd Ed., Princeton U. Press, Princeton, N.J., 1946.