MINERÍA DE SERIES TEMPORALES Y FLUJO DE DATOS

MÁSTER EN CIENCIA DE DATOS E INGENIERIA DE COMPUTADORES

Trabajo autónomo I: Series Temporales

Autores
José Ángel Díaz García
joseangeldiazg02@correo.ugr.es

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, Abril de 2018

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En este documento encontramos el resultado final alcanzado durante el estudio del apartado de series temporales, enmarcado dentro de la asignatura de 'Minería de Series Temporales y Flujos de Datos' del máster en Ciencia de Datos de la Universidad de Granada. En este primer capítulo, veremos una introducción al problema a resolver así como a los objetivos a alcanzar con esta práctica y la organización del trabajo.

1.1. Problema a resolver

Las series temporales han sido ampliamente estudiadas en la literatura y sus aplicaciones al mundo real son amplias y muy variadas. Estas aplicaciones, pueden ir desde temas financieros a problemas de stock o marketing, sin olvidarnos las ciencias de la tierra, entre las cuales destacan notablemente en problemas de climatología.

En esta práctica nos enfrentaremos por tanto a un problema de climatología en el que partiendo de series de datos de estaciones meteorológicas [1] tendremos que predecir los siguientes valores:

1. Valores de temperatura máxima en la primera semana de marzo a nivel de días. Es decir, 7 valores.
2. Valor de temperatura máxima en el mes de marzo a nivel de mes. Es decir, un solo valor.

Estaremos por tanto ante un **problema real de ciencia de datos** que tendrá nos enfrentará a dos escenarios similares pero que conllevan diferencias significativas en el procesado y las cuales, deberían ser solventadas por el hipotético analista de datos.

Estación meteorológica

La estación meteorológica objetivo está en la localidad de Coria en la provincia de Cáceres, localidad natural del autor del trabajo y aunque se conoce bien el lugar y su orografía nunca está de más cierta información que pueda ayudar en el proceso de análisis por ello, se han obtenido ciertos datos de la estación los cuales pueden verse en la tabla 1.1.

ID	*3526X*
PROVINCIA	*CÁCERES*
LATITUD	*400013N*
LONGITUD	*063327W*
ALTITUD	*313*

Tabla 1.1: Especificaciones de la estación meteorológica.

1.2. Objetivos

Los objetivos de esta practica podrían resumirse en los siguientes:

- Asentar y comprender la materia teórica de las series temporales vista durante el transcurso de la asignatura.
- Pre-procesar los datos de manera correcta para poder ser procesados por los modelos de series temporales.
- Usar correctamente los test estadísticos y comprender su potencia de cara a problemas de ciencia de datos.
- Obtener un modelo de predicción aceptable para ambos problemas descritos en la sección 1.1.

1.3. Organización del trabajo

La organización del presente documento, se centra en detallar cada uno de los pasos seguidos durante el estudio y resolución del problema planteado en esta introducción, tras la cual en el capítulo 2 veremos las explicaciones asociadas al preprocesado de datos, más concretamente a la limpieza de valores perdidos. El grueso de la memoria está en el capítulo 3 donde veremos los resultados obtenidos en la práctica así como el proceso para alcanzarlos. Finalizaremos la memoria con las conclusiones obtenidas en el transcurso de finalización de la misma en el capítulo 4.

Capítulo 2

Preprocesado

El preprocesado de datos es una de las tareas más importantes a las que un científico de datos debe enfrentarse. Esta tarea es de vital importancia y puede desembocar en el éxito o el fracaso de los modelos predictivos que se generen sobre los datos. En este capítulo veremos las técnicas de preprocesado aplicadas sobre nuestros datos.

2.1. Integración

El primer paso es la lectura de datos. Estos vienen en formato texto, por lo que para una correcta lectura deberemos pasar los factores y strings que correspondan a numéricos ya que de otro modo los posibles valores perdidos se etiquetaran como **cadenas vacías** y podrán repercutir en error. Nos quedamos por tanto en primera instancia con todos los datos numéricos y la fecha de su captación. El motivo de mantener todos los numéricos es mejorar el proceso siguiente de imputación de valores perdidos.

2.2. Valores perdidos

Para la imputación de valores perdidos primero usando la función *sum(is.na)* obtenemos el número de valores perdidos de nuestro variable Tmax, el resultado es 83, por lo que es un número muy a tener en cuenta y que deberá

ser predicho. Para la predicción de los valores perdidos nos apoyamos en los demás datos numéricos para hacer regresiones y predicciones, antes de ello, usamos un gráfico de patrón de perdidos para ver si tenemos alguna idea de que otras variables podremos usar para su predicción. El resultado de este gráfico puede verse en la figura 2.1.

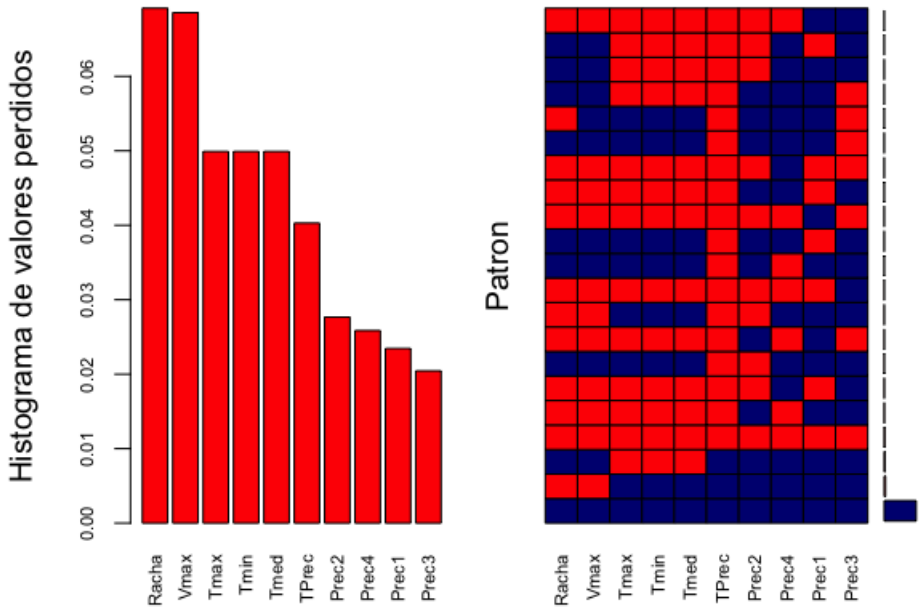

Figura 2.1: Patrón e histograma de valores perdidos.

Acorde a los resultados de la figura anterior, no podemos concluir que haya un patrón claro pero si podemos asegurar que cuando Tmax tiene un valor perdido las demás también lo tendrán con probabilidad por lo que utilizar regresiones será quizá mala solución, por ello nos decantaremos por el método **pmm** del paquete MICE. Tras lo cual, tendremos el dataset con una cantidad 0 de valores perdidos en Tmax.

2.3. Agregación

Este paso del preprocesado es solo necesario para la solución del problema mensual. Para ello, usaremos el comando aggregate y la función max para agregar para cada mes de cada año la mayor temperatura registrada de manera que podamos realizar predicciones y modelado de serie a nivel mensual. Para poder realizar esto, previamente hemos pasado la fecha de tipo string a tipo date y tras ello, hemos eliminado el día de la fecha de manera que podemos agregar por las parejas iguales de mes y año.

Capítulo 3

Análisis de las series

3.1. Serie Mensual

3.1.1. Análisis de tendencia y de estacionalidad

3.1.2. Estacionareidad

3.1.3. Modelado y predicción

3.2. Serie Diaria

3.2.1. Análisis de tendencia y de estacionalidad

3.2.2. Estacionareidad

3.2.3. Modelado y predicción

Capítulo 4

Conclusión

Tras la realización de la practica hemos podido constatar la potencia de los métodos disponibles en R para el modelado y predicción de series temporales. La curva de aprendizaje no sería muy elevada, al menos, para un control básico de las herramientas y test disponibles. Por otro lado, constatar el potencial del estos modelos y las innumerables aplicaciones del mundo real y las cuales un analista de datos debería conocer al menos en cierta medida.

Al margen de valoraciones personales, si nos centramos en los objetivos descritos al inicio de la memoria, podemos concluir que hemos cumplido con todos. Por un lado, con las explicaciones teóricas dadas en el código y durante la memoria podríamos dar por estudiado y asentado todo el contenido referente a series temporales, y aún mas sobre test estadísticos y su aplicación en ciencia de datos los cuales han quedado constatada su potencia al discernir entre que modelo se comportaría mejor y posterior mente comprobarlo con test reales. El pre-procesado de datos también se ha llevado a cabo de manera optima y bien podría mejorarse aún más realizando filtrados o algún tipo de selección que evitará varianzas elevadas. Finalmente, podemos concluir que los modelos generados son aceptables.

Bibliografía

[1] Website de la Agencia Estatal de Meteorología http://www.aemet.es/es/portada