Modelización de territorios – Grupo Electomedia

Uno de los objetivos iniciales de la práctica fin del bootcamp era la posibilidad de modelizar electoralmente un territorio amplio, pongamos una provincia, o incluso España entera, en función de una serie de territorios más pequeños, que podríamos suponer las secciones electorales. Creemos que hemos encontrado una manera relativamente sencilla de hacerlo, utilizando modelos de regresión lineal con o sin uso del método de PCA.

Hemos desarrollado esta parte de la práctica en dos cuadernos:

- **Modelación_territorio_secciones_LinReg_ZRZ**, en el que modelizamos mediante regresión lineal la provincia de Zaragoza para las elecciones de noviembre de 2019, en base a un número muy reducido de secciones de la misma provincia (menos de 10), y luego utilizamos el modelo para estimar el voto en 2016.

- **Modelación_territorio_secciones_PCA_Esp,** modelizando aquí toda España, mediante un pipeline con regresión lineal y PCA, en base a unas 66 secciones elegidas de nueve municipios, usando el modelo para estimar también el resultado en 2016.

En ambos casos el número de secciones elegidas es un porcentaje muy reducido, menos de 1% y 0,2%, respectivamente, del total de cada territorio modelizado.

Regresión lineal sin utilizar PCA

Una de las incógnitas es la selección de las secciones que utilizaremos para modelizar. En esencia se trata de un proceso de reducción de dimensionalidad, o de selección de características. Ya que los parámetros que definen una elección son los votos a los distintos partidos, no tenemos más de 25-30 filas de datos; las columnas son todas las secciones, mucho más numerosas.

Nos hemos basado en que uno de los métodos para reducir dimensiones es el eliminar aquellas columnas muy correlacionadas entre sí. Lo que hacemos es, partiendo del territorio utilizado para modelizar, en nuestros dos casos la provincia de Zaragoza y 9 municipios españoles, ir eliminando las secciones que tengan una alta correlación entre ellas. Fijamos un umbral que no podrán sobrepasar en su correlación con las demás, por lo que nos quedamos con las secciones (columnas) que de verdad aporten información.

En el caso de Zaragoza fijamos un umbral exigente, lo que hace que encontremos apenas 7 secciones, que las utilizamos para modelizar mediane una simple regresión lineal toda la provincia, compuesta por 880 secciones. Los resultados para noviembre de 2019 son muy satisfactorios.

A continuación, utilizando el mismo modelo para predecir los resultados de Zaragoza en 2016, tomado esta vez las secciones equivalentes en esa elección a las 7 seleccionadas en 2019. La estimación de voto se acerca también bastante al resultado real.

Regresión lineal utilizando PCA

Para el caso de España la metodología ha de ser algo diferente. La razón es que se deben de tomar secciones de muchas CCAA, entre ellas las que tienen partidos nacionalistas o regionalistas. No resulta práctico calcular una matriz de correlaciones con las 36.000 secciones de todo el país, por lo que elegimos una serie de municipios pequeños que representen las CCAA nacionalistas, y algunas que no lo son.

Un problema adicional es que, al ser apenas 25-30 datos a estimar, no podemos utilizar un número semejante de columnas, pues estaríamos, si su número fuese superior, en ante un sistema de ecuaciones incompatible, o si fuese algo menor, ante un peligro claro de overfitting. Por otro lado, al ser el voto en las CCAA tan variado, sí que nos hace falta contar con la información de muchas secciones.

Esta dicotomía pensamos que se soluciona mediante el uso de la reducción de dimensionalidad de la PCA, que es lo que utilizamos para modelizar toda España. Partiendo de las 283 secciones de los 9 municipios, ponemos un umbral de correlación relativamente bajo, seleccionando 66 secciones, que pasamos a meter un pipeline con una regresión lineal y un PCA de 10 dimensiones. El modelo para 2019 resulta ser bastante bueno, y generaliza bien cuando lo probamos al estimar el voto en toda España en 2016.

Los detalles del proceso de modelización se pueden consultar en los respectivos cuadernos.

Conclusiones

Por todo ello, creemos que tenemos un método de modelización bastante adecuado en general. Hemos probado a modelizar una provincia, como Zaragoza, usando secciones de otra, la de Burgos. En este caso comprobamos que, si bien el ajuste para 2019 era muy bueno, la generalización para 2016 resultó ser deficiente; la mejora de este proceso podría ser un siguiente paso a investigar.