Práctica 3

Ajuste de datos usando modelos lineales

27 de mayo de 2021

Aprendizaje Automático

FRANCISCO JAVIER SÁEZ MALDONADO

fjaviersaezm@correo.ugr.es

Índice

Introducción

En esta práctica, trataremos de realizar un estudio completo de un problema en el que se nos presenta un conjunto de datos y nuestro objetivo es seleccionar el mejor predictor lineal para este conjunto de datos dado. Concretamente, estudiaremos dos conjuntos de datos extraidos de la web UCI-Machine Learning Repository.

Utilizaremos uno de ellos para tratar de ajustar un modelo lineal a un problema de regresión, y otro conjunto diferente de datos para ajustar otro modelo lineal a un problema de clasificación multiclase. El objetivo será realizar un estudio de los datos, evitando en todo momento el *data snooping*, y argumentar si se utilizan ciertas técnicas de preprocesado de datos antes de escoger el modelo final.

Trataremos primero el problema de regresión y posteriormente el de clasificación.

1. Regresión

1.1. Estudio del conjunto de datos. Identificación de $\mathcal{X}, \mathcal{Y}, f$.

Lo primero que debemos hacer es realizar una buena comprensión de la información que tenemos sobre los datos para comprender un poco más nuestro problema.

Nuestro primer conjunto de datos, [2], contiene características de ciertos elementos superconductores. Junto con estas características, se nos presenta una *temperatura crítica*, que en el artículo en el que se estudia el conjunto de datos de manera más profunda [1] lo denominan T_c, obtenida para un superconductor que posea estas características. También se nos presenta un archivo en el que se nos dan las fórmulas químicas de los superconductores, pero este archivo no será relevante para nosotros.

Lo primero que nos encontramos acerca de nuestros datos es la siguiente tabla:

Características	Multivariable	Número de instancias	21263
Tipo de características	Reales	Número de atributos	81
Tareas asociadas	Regresión	Valores perdidos	$N\backslash A$

Tabla 1: Datos contenidos en el conjunto de datos Superconductivity.

Esta información nos resulta muy útil, pues obtenemos podemos observar que tenemos 81 atributos para cada una de las 21263 instancias. De aquí podemos obtener que el tamaño del conjunto de datos es bastante amplio, por lo que tendremos un buen conjunto de entrenamiento. Las características que obtenemos son reales, es decir, $x_i \in \mathbb{R}^{81}$. Para completar, vemos que no tenemos valores perdidos, por lo que nos ahorraremos en este caso tener que establecer una técnica para reconstruir estos valores.

Con la información proporcionada podemos decir que:

1. Nuestro conjunto de datos de entrenamiento será
$$\mathcal{X} = \{X_i \in \mathbb{R}^{81}, \text{ con } i = 1, \dots, 21263\}.$$
2. Nuestro conjunto de etiquetas, puesto que no se nos indica ninguna restricción sobre las temperaturas, podemos asumir que es:
$$\mathcal{Y} = \{y_i \in \mathbb{R}, \text{ con } i = 1, \dots, 21263\}.$$
3. Por último, nuestra función $f : \mathcal{X} \to \mathcal{Y}$ que asigne a cada vector de características una temperatura crítica.

Hay que anunciar que los siguientes gráficos de visualizado de datos se han realizado posteriormente a realizar la separación en conjuntos de *train* y *test* de nuestro conjunto de datos, para evitar en todo momento el *data snooping*.

Dibujamos ahora un gráfico en el que mostramos el diagrama de caja de los posibles valores que toma la temperatura T_c:

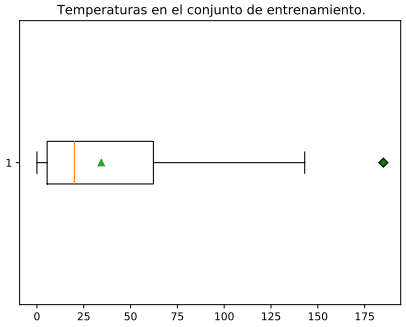

Figura 1: Diagrama de caja de las temperaturas T_c en el conjunto de entrenamiento.

Como podemos ver, aunque tenemos variabilidad en los valores de f, tenemos la mayoría de estos concentrados en el intervalo $[0, 50]$, lo cual nos indica que hay probabilidad de que los datos no sean separables o tengamos ruido en nuestra muestra. Como podemos ver, tenemos un dato que se aleja mucho de 1.5 por el rango intercuartílico, que es lo que representan los *bigotes* del diagrama de caja. Es por ello que podemos decir que este punto es posiblemente un *outlier*.

En cuanto a los valores de nuestros datos, si tomamos el primer elemento y calculamos la desviación típica σ y el resultado que obtenemos es:

```
Standard deviation of the first element of the dataset 2747.992
```

Por lo que obtenemos que claramente los valores de los diferentes atributos no están en el mismo rango de escala. Es por ello que previamente al entrenamiento realizaremos una estandarización por atributos de nuestro conjunto de datos.

Referencias

[1] Kam Hamidieh. "A Data-Driven Statistical Model for Predicting the Critical Temperature of a Superconductor". en. En: *arXiv:1803.10260 [stat]* (oct. de 2018). arXiv: 1803.10260. URL: http://arxiv.org/abs/1803.10260 (visitado 25-05-2021).

[2] *Superconductivty Data Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data (visitado 12-10-2018).