Aprendizaje automático

Proyecto final

Johanna Capote Robayna
Guillermo Galindo Ortuño
5 del Doble Grado en Informática y Matemáticas
Grupo A

UNIVERSIDAD
DE GRANADA

Índice

1. Definición del problema a resolver y enfoque elegido

Inicialmente se nos plantea un problema de regresión, cuyo objetivo es predecir la popularidad (contada en número de artículos compartidos en redes sociales) a partir de las características sobre los artículos publicados. Sin embargo, siguiendo las recomendaciones de los creadores del dataset trataremos este problema como un problema de clasificación binario, considerando todos aquellos valores menores e iguales que 1400 *shares* como una clase y los mayores como la otra.

El dataset consta de 60 atributos, dos de ellos no predictivos (*url* y *timedelta*), por lo tanto nuestro vector de características tendrá una longitud de 58. Por lo tanto identificamos los datos del problema como:

- $X : \mathbb{R}^{58}$
- $Y : \{-1, 1\}$
- $f : X \to Y$

2. Argumentos a favor de la elección de los modelos

3. Codificación de los datos de entrada par hacerlo útiles a los algoritmos

4. Valoración del interés de las variables para el problema y selección de un subconjunto

5. Normalización de las variables

6. Justificación de la función de pérdida usada

Como métrica de error utilizaremos el *accuracy*, la usual en este tipo de problemas. Esta medida expresa el error como un valor entre 0 y 1, siendo 0 cuando todos los puntos están bien clasificados y 1 cuando están todos mal clasificados. Para calcularla, dado un $h \in H$ el error viene dado por:

$$E_{in}(h) = \frac{1}{N} \sum_{x_n \in X} [[h(x) \neq y_n]]$$

Para visualizar y analizar el error utilizamos la matriz de confusión, aunque está no es una medida métrica, la mayoría de métricas se basan en esta matriz. Esta matriz es un método visual en el que podemos ver el rendimiento de un modelo supervisado. Esta matriz muestra los falsos positivos y los verdaderos positivos.

7. Selección del modelo lineal paramétrico y valoración de su idoneidad frente a otras alternativas

8. Aplicación de técnicas

9. Función de regularización

10. Valoración de los resultados

11. Justificación