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

El problema que inicialmente se nos plantea es del estimar la popularidad de un artículo (medido como número de veces que este es compartido) basándonos en una serie de características de este, como por ejemplo la longitud o si trata de temas como tecnología, estilo de vida, etc.

Aunque lo natural sería haberlo plantearlo como un problema de regresión, en nuestro caso hemos decidido enfocarlo como un problema de clasificación binario. Esto lo hemos hecho para poder utilizar y analizar modelos de clasficación tal y como hemos estudiado, que creemos que será más interesante. Siguiendo las recomendaciones de los creadores de la base de datos, trataremos este problema como un problema de clasificación binaria, considerando todos aquellos valores del atributo objetivo menores o iguales que un umbral (1400 en particular) como una clase y los mayores como la otra. Esto podemos interpretarlo como que queremos conocer si un artículo será popular o no (supera o no el umbral de *shares*).

El *dataset* consta de 61 atributos, siendo dos de ellos no predictivos (*url* y *timedelta*) y otro distinto el objetivo. Por tanto nuestro vector de características tendrá será de tamaño 58. Formalmente:

- Nuestro espacio muestral será $\mathcal{X} = \mathbb{R}^{58}$.
- El espacio de etiquetas será $\mathcal{Y} : \{-1, 1\}$.
- Nuestro objetivo será encontrar $f : X \to Y$ que estime si un artículo será popular o no (1 ó -1).

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