Regresión Lineal

- Entrenar la mejor línea a través de todos los puntos de datos. Este algoritmo puede ser :
    - Regresión Lineal Simple, en donde se realiza la predicción con una sola variable,
    - Regresión Lineal Múltiple, en donde se crea un modelo para la relación entre múltiples variables de entradas independientes o,
    - Regresión Polinomial en el que el modelo se convierte en una combinación no lineal de las variables características.

Regresión Lineal

VENTAJAS	DESVENTAJAS	SON ÚTILES
Fácil de entender y explicar	No se puede modelar relaciones complejas	Dar un vistazo al conjunto de datos
Rápido de modelar y es útil cuando la relación a modelar no es compleja y no tiene mucha información	No se pueden capturar relaciones no lineales sin transformar la entrada	Se tiene datos numéricos con muchas características
Menos propenso al sobreajuste	Puede sufrir con valores atípicos	Realizar predicción econométricas y respuestas de marketing

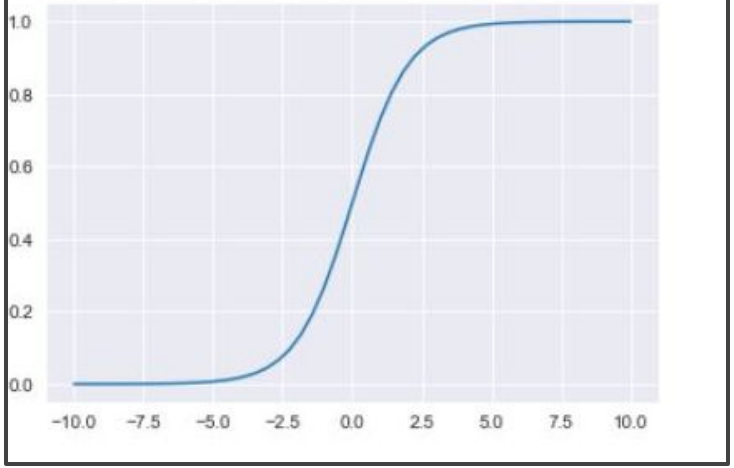
Regresión Logística

- El modelo de regresión logística se utiliza en escenarios de clasificación. Este algoritmo también se conoce como regresión logit, o clasificador de máxima entropía, y se encuadra dentro de los conocidos como modelos lineales generalizados.

- La regresión logística describe la probabilidad de que la variable objetivo pertenezca a una clase o a otra: dado un cierto valor límite, si el valor de la variable objetivo iguala o excede dicho valor, se devuelve como predicción la clase "positiva". En otro caso se devuelve como predicción la clase "negativa". Para ello se utiliza la función logística, que siempre devuelve un valor entre 0 y 1:

Regresión Logística

$$f(x) = \frac{1}{1 + e^{-x}}$$



Regresión Logística

- Scikit-Learn implementa la regresión logística en la ***clase sklearn.linear_model.LogisticRegression***.
  El argumento solver determina el algoritmo de optimización a usar ("newton-cg", "lbfgs", "liblinear", "sag" o "saga") y, en el caso de estar en un escenario multiclase, el argumento multi_class permite escoger la estrategia a seguir: uno contra los demás ("ovr", one-vs-rest) o un criterio de pérdida de entropía cruzada.

- ¿Cómo se entrena un dataset con una regresión logística?

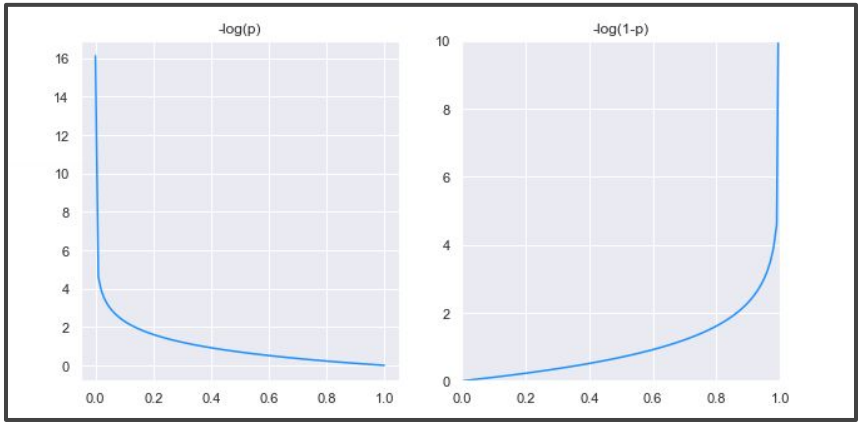
  Al igual que ocurre con un modelo de regresión lineal, la regresión logística también calcula una suma ponderada de las características predictivas, pero en lugar de devolver dicho resultado, lo pasa por la función sigmoide para devolver una probabilidad.

  A la hora de entrenar el modelo, el objetivo es asignar altas probabilidades a las muestras positivas (aquellas para las que la variable objetivo toma el valor 1) y bajas probabilidades para las muestras negativas, lo que se consigue con la siguiente función de coste para una única instancia:

$$c(\theta) = \begin{cases} -\log(\hat{p}) & si\ y = 1 \\ -\log(1-\hat{p}) & si\ y = 0 \end{cases}$$

Regresión Logística

- Se muestran a continuación ambas gráficas:



- Si a una muestra positiva se le asigna una probabilidad próxima a cero, el coste tiende a infinito (y, al revés, si se le asigna una probabilidad de uno, el coste es de cero). Por el contrario, si se trata de una muestra negativa, el asignarle una probabilidad próxima a uno supondría un coste también tendiente a infinito.

Regresión Logística

- La función de error correspondiente al dataset completo es simplemente el valor medio de la suma del error de todas las muestras, es decir:

$$C(\theta) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i))$$

Esta función se conoce como ***log loss***. Al contrario de lo que veíamos con la regresión lineal, no existe una "ecuación normal" que nos devuelva el valor mínimo de los argumentos. Sin embargo, esta función es convexa y derivable, lo que nos permite aplicar optimizadores como Gradient Descent para obtener el mínimo buscado.

Regresión Logística

En resumen, la regresión logística sigue estos pasos:

- Transformar la variable Y a predecir (que solo puede tomar dos valores: 0,1) en la probabilidad de Y=1, (que es igual a P) usando la distribución binomial de bernoulli.

- Calcular el Odds Ratio de la probabilidad P, que es igual a P / (1-P) = Odds Ratio

- Calcula el logaritmo natural del Odds Ratio, es decir: hace el ln(P/(1-P)) ó ln(Odds Ratio). Esto también puede ser interpretado como la cantidad de éxitos dividido entre la cantidad de fracasos. Se utiliza logaritmo natural para luego poder hacer la inversa, que es el exponencial. El logaritmo del Odds Ratio ahora es la variable independiente a predecir, también llamado logit.

- Estima la combinación lineal, o regresión lineal de las variables independientes, de la siguiente forma: ln(Odds Ratio) = $\beta 0 + \beta 1 * X1 + \beta 2 * X2$. Al resultado obtenido también le dicen βx.

- Se transforma ln(P/(1-P)) en $e^{\beta x} / (1+ e^{\beta x})$. Esto también lo representan como: $1 / (1+ e^{-\beta x})$ donde e es la constante matemática igual a 2.718281828

- Se estiman los coeficientes (es decir los β) de la ecuación anterior, usando el método de máxima verosimilitud (ver referencia Nro. 4)

Bibliografía

- https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
- The Elements of Statistical Learning | Trevor Hastie | Springer
- An Introduction to Statistical Learning | Gareth James | Springer
- Deep Learning | Ian Goodfellow | https://www.deeplearningbook.org/