Aprendizaje Supervisado

El objetivo del Machine Learning es crear un modelo predictivo en el cual cada vez que tenemos una nueva observación, el modelo sea capaz de predecir lo que queramos predecir para esa observación.

Para ello, tenemos que entrenar al modelo. Esto significa que tendremos un conjunto de observaciones (dataset) y hemos elegido un algoritmo de aprendizaje que produce el modelo. Para este modelo, veamos el siguiente diagrama:

Imagen de la pantalla de un computador

Descripción generada automáticamente con confianza baja

Por *supervisado* se refiere a que se conocen los valores de la columna objetivo (naranja) y la variables independientes (columna azul oscuro), y se llegan a dar los valores de la fila de valiables independiente (columna azul claro) para predecir el valor de la fila y columna objetivo (naranja claro).

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteRegresiones Lineales

Imagínese que se tienen los datos como se ve en la figura de arriba, se puede observar que la nube de puntos aparentemente tiene una tendencia a parecerse a una línea recta. Este tipo de situaciones se conoce como regresión lineal.

Supongamos que tenemos una tabla que contiene mediciones, . Es decir, hay características y queremos explicar en función de las otras variables. Esta explicación se describe a través de

El objetivo del método es buscar los valores de tal que minimicen el error

Para cuando n=1, se puede obtener una gráfica de la forma

Imagen que contiene Diagrama

Descripción generada automáticamente

Ahora veamos los tipos de errores que se pueden utilizar para evaluar la eficacia del método

**Error absoluto medio**

Es decir, la media de las diferencias entre la variable objetivo y las predicciones sin el signo.

MAE es una métrica robusta, en cuanto a que no varía mucho si hay valores extremos en los datos. El error se puede interpretar como unidades de la variable objetivo (por ejemplo, si la variable objetivo es en dólares MAE estará también en dólares).

**Error cuadrático medio**

Dado que el MSE se define en unidades al cuadrado, lo cual no es intuitivo (¿dolares cuadrados?), generalmente se usa su raíz.

**Raíz del error cuadrático medio**

La Raíz del Error Cuadrático Medio (Root Mean Squared Error o RMSE) se diferencia del MSE en que el resultado se puede medir en las mismas unidades que la variable objetivo

Sin embargo, tiene un problema y es que da más importancia a los errores grandes.

**Coeficiente de determinación**

El coeficiente de determinación (R2 o R-squared) mide la porción de la varianza de la variable objetivo que se puede explicar por el modelo.

R2 tiene un valor máximo de 1 (cuando el modelo explica toda la varianza) , aunque puede tener valores negativos.

Hay varias formas de definir R2 , pero una de las más sencillas es simplemente la correlacion (definida como la Correlación de Pearson) entre la variable objetivo y las predicciones , elevada al cuadrado.

Un problema importante que tiene R2 es que no nos indica si el modelo explica la varianza debido a que está sobreajustado (overfitted). Por eso una medida mejor es el Coeficiente de Determinacion ajustado (Adjusted R-squared), que tiene en consideracion la complejidad del modelo.