Cuestiones y terminología de la Ciencia de Datos

Christopher González Delgado

1. Cuestiones y Términos I

1.1. ¿Qué es la ciencia de datos?

La ciencia de datos se encarga de hacer predicciones precisas y de automatizar transacciones. Esta se centra más en los propios datos y en la construcción de nuevos sistemas capaces de procesarlos. Algunos de los campos que se abarcan son la visualización de datos, ingeniería de datos, pre-procesamiento de los datos o la toma de decisiones basadas en datos.

1.2. ¿Diferencias entre aprendizaje supervisado, no supervisado y reforzado?

Aprendizaje Supervisado: Los modelos de aprendizaje supervisado aprenden a partir de un conjunto de datos etiquetado para generar predicciones al introducir nuevos datos.
Aprendizaje No Supervisado: En un modelo sin supervisión los datos no están ordenados. El algoritmo tratará de buscar patrones o ideas por sí mismos.
Aprendizaje Reforzado: Este aprendizaje depende del agente de aprendizaje y en él se implementa la psicología conductista para determinar la mejor solución posible.

1.3. Regresión Lineal

Es un modelo matemático que establece una relación entre una variable dependiente Y y una variable independiente X mediante la búsqueda de dos parámetros (a,b)que consigan ajustar mejor estos datos mediante una recta $Y = aX + b$. Aquí a es la pendiente y b la intersección.

1.4. Regularización. Regularización L1 y L2

En algunas ocasiones nuestro modelo sufrirá overfitting. Esta situación ocurre cuando el modelo ha sido entrenado y funciona perfectamente en el set de entrenamiento, pero su funcionamiento en el set de testeo es muy pobre. Al introducir la regularización estamos tratando de eliminar el overfitting.

Regularización L1: También conocida como regularización Lasso, introduce un hiperparámetro λ que se encarga de reducir el valor de los coeficientes de las características menos importantes a cero. Esta corrección se añade a la función de coste como:

$$L1 = \lambda \sum_i |W_i| \tag{1}$$

Regularización L2: La regularización L2 o regularización Ridge es similar a la regularización Lasso, salvo que en este caso las características se elevan al cuadrado.

$$L2 = \lambda \sum_i |W_i|^2 \tag{2}$$

1.5. ¿Qué es R^2?

Se trata de una medida estadística que nos indica cómo de cerca están los datos de la línea de ajuste. Generalmente, al tener un mayor R^2 el modelo se ajustará mejor a los datos.

$$R^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3}$$

Donde y_i son los datos reales, $\hat{y}_i$ es la predicción y $\hat{y}$ la media de los valores reales.
Hay un problema con este coeficiente y es que siempre crecerá al aumentar el número de características o predictores, por lo que se debe definir un R^2 ajustado, R^2_{adj}

$$R^2_{adj} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} \tag{4}$$

En esta ecuación, N es el tamaño de la muestra y p es el número de predictores utilizados.

1.6. Error cuadrático medio

El error cuadrático medio (ECM) nos dice que tan cerca está una línea de regresión del conjunto de puntos. Se toman las distancias de los puntos a las líneas y se elevan al cuadrado.

$$ECM = \frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2 \tag{5}$$

1.7. Regresión Logística

De forma similar a la regresión lineal, es un modelo matemático que nos permite hacer predicciones a partir de ciertas características pero en esta ocasión la predicción será de una variable categórica. Es decir, cuando el resultado no es un número sino una categoría determinada entra en juego la regresión logística.

1.8. Diferencias entre regresión lineal y logística

Tanto la regresión lineal como la logística son las formas más básicas de regresión que se conocen. La principal diferencia entre estas es que cuando queremos predecir valores continuos (precios, litros de agua consumidos, etc...) utilizaremos una regresión lineal. En el caso de que necesitemos un resultado categórico (animal o persona, un color, etc...) se usará la regresión logística.

1.9. Errores en Machine Learning: Bias y Varianza

Error de Bias: Es la diferencia entre los valores verdaderos y la predicción esperada de nuestro modelo. El error de bias se produce cuando los modelos son muy rígidos (e.g. modelos lineales) que no son capaces de comprender los datos (e.g. ajuste lineal de datos con un patrón exponencial o polinómico).
Error de Varianza: Este error se introduce como consecuencia de la variabilidad de la predicción de un modelo para un determinado punto. Si entrenamos un modelo con diferentes sets de entrenamiento observaremos que el resultado que ofrece el modelo para un mismo punto a la hora de testearlo es diferente.

1.10. Intercambio Bias-Varianza

Cuando entrenamos modelos de predicción estamos buscando que tengan tanto bias como varianza baja, pero resulta que los dos errores que se han introducido anteriormente están totalmente ligados entre sí. Existe por tanto un intercambio o una compensación entre bias-varianza. Es decir, siempre que provoquemos una disminución en el error de bias, aumentará el error de varianza, y viceversa. Por tanto, debemos buscar el punto de equilibrio que nos genere el modelo más eficiente.