Aprendizaje Supervisado
Regresión Lineal

Profesor: Juan Gamarra Moreno

Introducción

Aprendizaje Supervisado

- **Aprendizaje Supervisado** es una subcategoría del aprendizaje automático (Machine Learning) en la cual un modelo se entrena utilizando un conjunto de datos etiquetados. En este contexto, "etiquetado" significa que cada ejemplo de entrenamiento contiene una entrada (características o features) y una salida deseada (etiqueta o target). El objetivo del aprendizaje supervisado es que el modelo aprenda a mapear las entradas a las salidas de manera precisa, de forma que, cuando se le presenten nuevos datos no vistos, pueda predecir correctamente la salida correspondiente.

Aprendizaje Supervisado

El proceso implica los siguientes pasos:

1. **Entrenamiento**: El modelo recibe un conjunto de datos etiquetados y ajusta sus parámetros internos para minimizar el error entre sus predicciones y las salidas reales.
2. **Validación**: Una vez entrenado, el modelo se evalúa con un conjunto de datos distintos para comprobar su capacidad de generalización, es decir, su habilidad para funcionar bien con nuevos datos no utilizados en el entrenamiento.
3. **Predicción**: El modelo entrenado se usa para hacer predicciones sobre nuevos datos desconocidos.

Algoritmos de Aprendizaje Supervisado

Ejemplos comunes de algoritmos de aprendizaje supervisado incluyen la regresión lineal, los árboles de decisión, las máquinas de soporte vectorial (SVM) y las redes neuronales.

El aprendizaje supervisado se divide en dos grandes categorías:

- **Clasificación**: Cuando la salida es discreta (como clasificar imágenes de perros y gatos).
- **Regresión**: Cuando la salida es continua (como predecir el precio de una casa en función de ciertas características).

Introducción a los Modelos de Regresión

Regresión Lineal

La **regresión lineal** es un método estadístico utilizado para modelar y analizar la relación entre una variable dependiente (o de salida) y una o más variables independientes (o predictoras). En esencia, la regresión lineal busca ajustar una línea (o plano en el caso de varias variables) que mejor describa cómo las variables independientes influyen en la variable dependiente.

- **Variable dependiente (Y)**: Es la variable que intentamos predecir o explicar.
- **Variables independientes (X)**: Son las variables que utilizamos para predecir o explicar la variable dependiente.

Regresión Lineal

La regresión lineal puede dividirse en dos tipos principales:

- **Regresión lineal simple**: Donde solo hay una variable independiente.
- **Regresión lineal múltiple**: Donde hay más de una variable independiente

Regresión Lineal Simple

$$y = \beta_0 + \beta_1 x + \epsilon$$

Donde:

- y es la variable dependiente (lo que queremos predecir),
- x es la variable independiente,
- β_0 es el intercepto o término constante (valor de y cuando x=0),
- β_1 es el coeficiente que mide el cambio en y por un cambio unitario en x,
- ϵ es el término de error, que captura la variabilidad no explicada por x.

Regresión Lineal Múltiple

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Donde:

- y es la variable dependiente (lo que queremos predecir),
- $x_1, x_2, \ldots, x_n$ son las variables independientes,
- $\beta_0, \beta_1, \ldots, \beta_n$ son los coeficientes que cuantifican la relación entre cada variable independiente y la variable dependiente,,
- ϵ es el término de error

Hipótesis y Suposiciones del Modelo de Regresión Lineal

Para que el modelo de regresión lineal sea válido y genere resultados confiables, se deben cumplir ciertas hipótesis y suposiciones:

- **Linealidad**: Existe una relación lineal entre la variable dependiente y las variables independientes. Esto significa que un cambio en una variable independiente se traduce en un cambio constante en la variable dependiente.

Hipótesis y Suposiciones del Modelo de Regresión Lineal

- **Homocedasticidad**: Los residuos (diferencias entre los valores predichos y observados) deben tener una varianza constante a lo largo de todos los niveles de las variables independientes. Si la varianza no es constante, se dice que hay **heterocedasticidad**, lo que puede invalidar el modelo.

Hipótesis y Suposiciones del Modelo de Regresión Lineal

- **Independencia de los errores**: Los errores del modelo (residuos) deben ser independientes entre sí, lo cual significa que no debe haber autocorrelación entre ellos.

Hipótesis y Suposiciones del Modelo de Regresión Lineal

- **Normalidad de los errores**: Los residuos deben seguir una distribución normal. Esto es importante para garantizar la validez de las pruebas estadísticas y la precisión de los intervalos de confianza.

Limitaciones de la Regresión Lineal

Aunque la regresión lineal es una herramienta poderosa, tiene varias limitaciones:

- **Sobreajuste**: Si se añaden demasiadas variables independientes, el modelo puede ajustarse demasiado a los datos de entrenamiento, capturando ruido en lugar de la verdadera relación subyacente. Esto genera un modelo que funciona bien en el entrenamiento, pero que generaliza mal para nuevos datos.

Limitaciones de la Regresión Lineal

- **Sensibilidad a los valores atípicos**: La regresión lineal es altamente sensible a los valores atípicos. Un solo valor fuera de lo común puede afectar drásticamente la pendiente de la línea de regresión, alterando las predicciones.

Limitaciones de la Regresión Lineal

- **Colinealidad**: En la regresión lineal múltiple, si dos o más variables independientes están altamente correlacionadas entre sí, puede ser difícil determinar qué variable está verdaderamente influyendo en la variable dependiente. Esto se denomina **multicolinealidad** y puede desestabilizar los coeficientes de la regresión, haciendo que el modelo sea menos interpretable.

Regularización

Regularización

- **Regularización** es una técnica en Machine Learning utilizada para prevenir el sobreajuste (overfitting) de un modelo. El sobreajuste ocurre cuando un modelo se ajusta demasiado bien a los datos de entrenamiento, capturando el ruido o variabilidad irrelevante, lo que reduce su capacidad de generalización a nuevos datos. La regularización introduce una penalización a los coeficientes del modelo para evitar que estos crezcan demasiado grandes y, con ello, reduce la complejidad del modelo.

Ridge Regression (L2 Regularization)

- En **Ridge**, se añade una penalización proporcional al cuadrado de los coeficientes (β) del modelo. Esta penalización evita que los coeficientes crezcan demasiado, lo que puede ser un síntoma de sobreajuste.
- La ecuación de costo para Ridge es:

$$C(\beta) = SSE + \lambda \sum_{i=1}^{n} \beta_i^2$$

Donde:

- *SSE* es la suma del error cuadrático.
- λ es el parámetro de regularización que controla la fuerza de la penalización.
- $\sum \beta_i^2$ es la suma de los cuadrados de los coeficientes.

- A medida que λ aumenta, los coeficientes se reducen más, lo que puede llevar a un mejor ajuste en datos nuevos.

Lasso Regression (L1 Regularization)

- En **Lasso**, se añade una penalización proporcional al valor absoluto de los coeficientes. Una característica importante de Lasso es que tiende a forzar a algunos coeficientes a ser exactamente cero, lo que también actúa como un método de selección de características.
- La ecuación de costo para Lasso es:

$$C(\beta) = SSE + \lambda \sum_{i=1}^{n} |\beta_i|$$

- A diferencia de Ridge, Lasso puede reducir a cero algunos coeficientes, eliminando así características irrelevantes del modelo, lo que lo hace útil cuando se tiene un conjunto de datos con muchas variables.

Elastic Net

- Es una combinación de Ridge y Lasso, que introduce tanto penalizaciones L1 como L2 en el modelo. Se usa cuando hay colinealidad entre las variables independientes y se busca beneficiarse de las propiedades de ambos métodos de regularización.

¿Por qué usar Regularización?

- **Prevenir el sobreajuste**: La regularización fuerza al modelo a encontrar una solución más simple y menos compleja, mejorando su capacidad de generalización.
- **Reducir la varianza**: Modelos con muchos coeficientes pueden ser muy variables y sensibles a cambios en los datos. La regularización ayuda a controlar esta variabilidad.
- **Mejorar la estabilidad del modelo**: En presencia de multicolinealidad o datos ruidosos, la regularización mejora la estabilidad de los coeficientes.

La clave está en ajustar el parámetro de regularización (λ) correctamente, equilibrando el sesgo y la varianza para lograr un buen rendimiento en datos no vistos.

Evaluación del Modelo de Regresión

Introducción

Al entrenar un modelo de regresión, es fundamental evaluar su desempeño para determinar qué tan bien puede predecir nuevos datos. A continuación, se describen:

- cuatro métricas de rendimiento comúnmente utilizadas y
- una técnica esencial para garantizar la validez del modelo.

Error Cuadrático Medio (MSE)

El **Error Cuadrático Medio** es una métrica que mide la precisión de un modelo de regresión calculando la diferencia entre los valores predichos y los valores reales.

La fórmula del MSE es:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Donde:

- y_i es el valor real de la variable dependiente.
- $\hat{y}_i$ es el valor predicho por el modelo.
- n es el número de observaciones.

Error Cuadrático Medio (MSE)

El MSE mide el promedio de los errores al cuadrado, por lo que:

- Un **MSE bajo** indica que las predicciones del modelo están cerca de los valores reales.
- Un **MSE alto** sugiere que el modelo está prediciendo de manera imprecisa.

El MSE es sensible a los valores atípicos, ya que, al elevar los errores al cuadrado, se amplifica el impacto de los grandes errores.

Error Absoluto Medio (MAE)

El **MAE** mide el promedio de la magnitud de los errores (es decir, las diferencias absolutas entre los valores predichos y los reales), lo que lo convierte en una métrica menos sensible a los valores atípicos que el MSE.

La fórmula del MAE es:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

El **MAE** ofrece una medida fácil de interpretar, ya que da una idea directa del promedio de los errores cometidos en las predicciones. A diferencia del MSE, no penaliza de manera desproporcionada los grandes errores, por lo que es más robusto en presencia de valores atípicos.

Raíz del Error Cuadrático Medio (RMSE)

El **RMSE** es simplemente la raíz cuadrada del MSE. Al tomar la raíz cuadrada, se devuelve la métrica a las mismas unidades que la variable dependiente, lo que puede facilitar la interpretación de la magnitud de los errores.

La fórmula del RMSE es:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

El RMSE tiene las mismas ventajas que el MSE, en términos de penalizar los grandes errores más que los pequeños. Sin embargo, al estar en las mismas unidades que la variable dependiente, es más intuitivo para medir la desviación promedio de las predicciones en comparación con los valores reales.

Coeficiente de Determinación (R^2)

El ***R²***, también conocido como **coeficiente de determinación**, mide qué proporción de la variación en la variable dependiente está explicada por las variables independientes en el modelo.

La fórmula de R^2 es:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Donde:

- $\bar{y}$ es el promedio de los valores reales.

Coeficiente de Determinación (R^2)

El valor de R^2 varía entre 0 y 1:

- $\boldsymbol{R^2 = 1}$: El modelo predice perfectamente todos los datos (todos los puntos caen sobre la línea de regresión).
- $\boldsymbol{R^2 = 0}$: El modelo no explica ninguna de las variaciones de los datos.
- $\boldsymbol{R^2}$ **negativo**: Esto indica que el modelo es peor que simplemente predecir la media de los datos.

Un $\mathbf{R^2}$ **alto** sugiere que el modelo explica bien las variaciones en los datos, mientras que un $\mathbf{R^2}$ **bajo** indica que el modelo no captura bien la relación entre las variables.

Validez del Modelo: Validación Cruzada

La **validación cruzada** es una técnica esencial para garantizar que el modelo generalice bien a nuevos datos, es decir, para evitar el sobreajuste (**overfitting**). La idea principal es dividir los datos en varios subconjuntos, entrenar el modelo en una parte de los datos, y probarlo en otra, para así obtener una medida más confiable de su rendimiento.

Validez del Modelo: Validación Cruzada

K-Fold Cross-Validation:

Es uno de los métodos más comunes de validación cruzada. El procedimiento es el siguiente:

1. Se divide el conjunto de datos en **k** subconjuntos de tamaño aproximadamente igual.
2. Se entrena el modelo usando **k-1** subconjuntos y se evalúa en el subconjunto restante.
3. Este proceso se repite **k** veces, cada vez usando un subconjunto diferente como conjunto de prueba.
4. Al final, se promedian los resultados obtenidos de cada iteración para obtener una métrica de rendimiento global.

Este proceso asegura que cada observación se utilice tanto para el entrenamiento como para la evaluación del modelo.

Validez del Modelo: Validación Cruzada

Importancia de la Validación Cruzada:

- **Evita el sobreajuste**: Al probar el modelo en diferentes subconjuntos, se asegura que el modelo no se ajuste demasiado a un solo conjunto de datos.
- **Generalización**: La validación cruzada ayuda a verificar si el modelo generaliza bien, es decir, si su rendimiento es consistente cuando se enfrenta a datos nuevos no vistos durante el entrenamiento.
- **Hiperparámetros**: Es útil para ajustar hiperparámetros, como el valor de regularización en Ridge y Lasso, asegurando que el modelo no solo funcione bien en los datos de entrenamiento, sino también en datos nuevos.

La combinación de métricas de rendimiento como el MSE y el R^2, junto con técnicas de validación cruzada, es clave para evaluar adecuadamente la calidad de un modelo de regresión y garantizar que sea capaz de predecir de manera precisa y consistente en diferentes conjuntos de datos.

Regresión Polinomial y Regresión con Variables Polinómicas

Regresión Polinomial

- La regresión polinomial es una extensión de la regresión lineal en la que se introducen términos polinomiales de las variables independientes para modelar relaciones no lineales entre la variable dependiente (objetivo) y las variables independientes. En otras palabras, mientras que en la regresión lineal la relación entre las variables es una línea recta, en la regresión polinomial, la relación puede ser una curva.

Características de la Regresión Polinomial

- **Forma del Modelo**: En la regresión polinomial, la relación entre la variable dependiente "y" y la variable independiente "x" se modela mediante un polinomio de grado n:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \in$$

Donde:

- y es la variable dependiente.
- $x, x^2, x^3, \ldots, x^n$ son las potencias de la variable independiente.
- $\beta_0, \beta_1, \ldots, \beta_n$ son los coeficientes que representan la magnitud de la influencia de cada término polinomial.
- ϵ es el término de error.

Características de la Regresión Polinomial

- **Relación no lineal**: Aunque el modelo utiliza potencias de las variables independientes, sigue siendo un **modelo lineal en los coeficientes** (es decir, las βs), lo que lo distingue de los modelos no lineales en los parámetros. Sin embargo, el comportamiento de la función predicha puede ser no lineal, lo que lo convierte en una técnica útil para capturar relaciones curvas entre las variables.

Características de la Regresión Polinomial

- **Grado del polinomio**: El grado del polinomio, *n*, determina la complejidad de la curva. Por ejemplo:
  - Un polinomio de **grado 1** es simplemente una línea recta (regresión lineal).
  - Un polinomio de **grado 2** genera una parábola.
  - Un polinomio de **grado 3** y superiores pueden generar curvas más complejas.

Ventajas de la Regresión Polinomial

Ventajas de la Regresión Polinomial:

- **Captura relaciones no lineales**: Es útil cuando los datos no pueden ser bien modelados con una línea recta, sino que requieren una curva para describir correctamente la relación entre las variables.
- **Simplicidad matemática**: A pesar de la complejidad de la curva, el modelo sigue siendo relativamente fácil de entrenar usando los mismos métodos que la regresión lineal, como los mínimos cuadrados

Desventajas de la Regresión Polinomial

Desventajas de la Regresión Polinomial:

- **Sobreajuste**: A medida que el grado del polinomio aumenta, el modelo puede volverse demasiado flexible, ajustándose demasiado bien a los datos de entrenamiento y capturando el ruido, lo que reduce su capacidad de generalización a nuevos datos.
- **Extrapolación peligrosa**: Fuera del rango de los datos de entrenamiento, las predicciones pueden ser erráticas, ya que los polinomios de alto grado tienden a tener comportamientos muy diferentes en regiones donde no hay datos.
- **Sensibilidad a los valores atípicos**: Los modelos polinomiales son muy sensibles a valores atípicos, ya que estos pueden cambiar drásticamente la forma de la curva.

Regresión con variables polinómicas

- La **regresión con variables polinómicas** es una técnica de regresión que extiende un modelo lineal añadiendo **términos polinómicos** de las variables independientes. Es útil cuando la relación entre la variable dependiente (resultado o predicción) y las variables independientes (predictoras) no es estrictamente lineal. En lugar de ajustar un modelo lineal simple, se incorporan términos elevados a diferentes potencias o productos entre las variables independientes para capturar esta no linealidad.

Regresión con variables polinómicas

Concepto Clave:

- En la regresión con variables polinómicas, no necesariamente se ajusta un modelo polinómico puro. En su lugar, se agregan términos polinómicos al conjunto de características (predictoras) y se ajusta un modelo, que podría ser lineal o de otro tipo, a este nuevo conjunto de variables.

Regresión con variables polinómicas

Ejemplo:

- Supongamos que tenemos una variable dependiente y y una variable independiente x, y creemos que la relación no es lineal. En lugar de usar solo x, podemos agregar términos como x^2 y x^3 para capturar la no linealidad. El modelo con variables polinómicas sería:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \in$$

- Aquí, estamos incluyendo términos polinómicos de x (es decir, x^2 y x^3) como **nuevas variables independientes**.

Regresión con variables polinómicas

Aplicación a Más de una Variable Independiente:

Si tenemos múltiples variables independientes, también podemos agregar términos polinómicos de estas variables, incluidos los productos cruzados entre ellas.

Por ejemplo, con dos variables x_1 y x_2, podríamos crear un modelo con variables polinómicas como este:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 (x_1 \times x_2) + \in$$

En este caso, hemos añadido:

- Términos polinómicos cuadráticos de x_1 y x_2 (x_1^2 y x_2^2).
- Un término de interacción entre las variables x_1 y x_2 $(x_1 \times x_2)$.

Este enfoque permite capturar relaciones no lineales y efectos de interacción entre las variables independientes.

Regresión con variables polinómicas

¿Por qué usar variables polinómicas?

- La principal razón para incluir **variables polinómicas** en un modelo de regresión es cuando se observa que una relación lineal entre las variables no captura bien los patrones en los datos, es decir, la relación entre la variable independiente y la dependiente parece curvilínea o no lineal.

Regresión con variables polinómicas

Pasos para implementar la Regresión con Variables Polinómicas:

1. **Transformar las variables independientes**: Crear nuevas variables que sean potencias o productos de las variables originales.
   - Ejemplo: Para una variable x, podrías incluir x^2, x^3, etc.
   - Para múltiples variables x_1, x_2, podrías incluir $x_1 \times x_2$, x_1^2 y x_2^2, y otros términos combinados.
2. **Ajustar el modelo**: Aplicar una técnica de regresión (por ejemplo, regresión lineal) al conjunto de datos transformado que incluye los términos polinómicos.
3. **Evaluar el modelo**: Usar métricas de rendimiento como el **MSE**, **MAE**, **RMSE**, o el $\mathbf{R^2}$ para evaluar qué tan bien el modelo con variables polinómicas ajusta los datos en comparación con un modelo lineal simple.

Regresión con variables polinómicas

Ventajas:

- **Captura de relaciones no lineales**: Los términos polinómicos permiten representar mejor las relaciones complejas entre las variables.
- **Flexibilidad**: Puede manejar relaciones tanto cuadráticas, cúbicas u otras no lineales.
- **Interacciones**: Permite modelar interacciones entre variables independientes al incluir productos entre ellas.

Regresión con variables polinómicas

Desventajas:

- **Sobreajuste**: A medida que se agregan más términos polinómicos, el modelo puede ajustarse demasiado bien a los datos de entrenamiento y no generalizar bien a nuevos datos.
- **Complejidad**: El modelo se vuelve más complejo y difícil de interpretar a medida que se agregan más variables y términos de mayor grado.
- **Multicolinealidad**: Los términos polinómicos pueden introducir multicolinealidad en el modelo, lo que puede hacer que los coeficientes sean inestables.