Trabajo Práctico
Matemática IV
Entrega Regresión Lineal

Augusto Conti, Nicolás Tenaglia, Sofía Celano

Septiembre 2024

Link al drive con archivos adjuntos:

`https://drive.google.com/drive/folders/1ZPGol9eUOeiuNtkBALlg9M-wN4_uLRFH?usp=sharing`

Introducción

En el primer ejercicio estimamos la recta de regresión simple tomando como característica más relevante al valor de la claúsula de recesión del jugador (release_clause_eur). Luego realizamos un test de hipótesis con un intervalo de confianza del 95% y lo comparamos con el intervalo de predicción de valores futuros con la misma confianza.
En el segundo ejercicio estimamos la recta de regresión múltiple tomando como características más relevantes al valor de la claúsula por recesión del jugador (nuevamente), al rendimiento promedio del jugador (overall) y al potencial (potential) del jugador. Ejecutamos el algoritmo de descenso del gradiente y lo analizamos y comparamos con lo obtenido anteriormente.
En el tercer ejercicio analizamos de forma teórica el comportamiento método mencionado anteriormente.

Desarrollo

1 Predicción del valor de mercado

a) Recta de regresión para predecir el valor de mercado de un jugador a partir de la característica más relevante (a la que se destinará mayor proporción del presupuesto), respaldada por:

- i) Prueba de significancia de regresión, coeficiente de determinación (R^2) y correlación lineal (r).

- ii) Inferencias sobre los parámetros de la recta, estimando las fluctuaciones con una confianza del 95%.
- iii) La proporción de veces que el valor de mercado supera la incertidumbre de predicción comparada con la respuesta media del valor de mercado para una característica fija, ambas con la misma confianza y ancho mínimo.

Coeficiente de determinación (R^2) El coeficiente de determinación R^2 mide qué tan bien el modelo se ajusta a los datos. Se puede calcular usando las sumas de cuadrados:

$$R^2 = 1 - \frac{SS\_res}{SS\_tot}$$

Donde:

- $SS\_res = \sum(y_i - \hat{y}_i)^2$ es la suma de los errores residuales.
- $SS\_tot = \sum(y_i - \bar{y})^2$ es la suma total de los cuadrados (variación total en y).

Correlación lineal (r)
La correlación de Pearson r se calcula con la fórmula:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Donde:

- $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$
- $S_{xx} = \sum(x_i - \bar{x})^2$
- $S_{yy} = \sum(y_i - \bar{y})^2$

Prueba de significancia de la regresión (valor p para la pendiente)
La prueba de significancia de la regresión implica usar el t-statistic para la pendiente b_1. El estadístico t se calcula como:

$$t = \frac{b_1}{SE(b_1)}$$

Donde $SE(b_1)$ es el error estándar de la pendiente y se puede calcular como:

$$SE(b_1) = \frac{\sqrt{SS_{res}/(n-2)}}{\sqrt{S_{xx}}}$$

Donde:

- $SS_{res} = \sum(y_i - \hat{y}_i)^2$ es la suma de residuos.

- n es el número de puntos de datos.

Modelo de regresión lineal
La regresión lineal se ajusta a una ecuación de la forma:

$$y = b_0 + b_1 x$$

Donde:

- b_0 es la intersección (ordenada al origen).
- b_1 es la pendiente.

Explicación:

Valores obtenidos:

Coeficiente de determinación ($\mathbf{R^2}$): 0.9880
Correlación lineal (r): 0.9940
t-statistic para la pendiente: 1216.2369

1. **Coeficiente de determinación ($\mathbf{R^2}$):** Mide la proporción de la variación total en y que es explicada por el modelo lineal. Va de 0 a 1, donde 1 indica que el modelo explica perfectamente los datos.
2. **Correlación lineal (r):** Mide la fuerza de la relación lineal entre x e y. Va de -1 a 1, donde 1 indica una correlación positiva perfecta.
3. **Prueba de significancia de la regresión (t-statistic):** Evalúa si la pendiente (b_1) es significativamente diferente de cero. Si el valor absoluto de t es grande, la pendiente es significativa.

Error estándar de los parámetros
Para obtener las inferencias sobre los parámetros, necesitamos calcular el error estándar de los coeficientes b_0 (ordenada al origen) y b_1 (pendiente). El error estándar se basa en el residuo del modelo.
Intervalos de confianza del 95%
El intervalo de confianza para un coeficiente de regresión b_i está dado por:

$$IC = b_i \pm t_{\alpha/2} \cdot SE(b_i)$$

Donde:

- $t_{\alpha/2}$ es el valor crítico de la distribución t de Student para un nivel de confianza del 95% y los grados de libertad correspondientes ($n - 2$ para una regresión lineal simple).

- $SE(b_i)$ es el error estándar del coeficiente.

Cálculo de la proporción de veces que el valor de mercado supera la incertidumbre de predicción

Para calcular la proporción de veces que el valor de mercado supera la incertidumbre de predicción comparada con la respuesta media del valor de mercado para una característica fija, usando la misma confianza y ancho mínimo, se deben considerar varios conceptos estadísticos y probabilísticos en un contexto de regresión lineal.

Este proceso involucra:

1. **Intervalo de predicción:** Estima el rango en el que es probable que caiga una nueva observación de y (el valor de mercado) para un valor fijo de x (la característica fija). Los intervalos de predicción son más amplios que los intervalos de confianza porque incluyen la incertidumbre en los nuevos datos.

2. **Intervalo de confianza:** Estima el rango en el que probablemente se encuentra el valor medio de y para un valor fijo de x, pero no incluye la variabilidad de los nuevos datos.

Estrategia:

1. **Intervalo de predicción:** Compara el valor de mercado real con el intervalo de predicción.

2. **Intervalo de confianza de la media:** Obtén el intervalo de confianza de la media para un valor específico de x.

3. **Proporción de veces que el valor de mercado supera la incertidumbre de predicción:** Esto se refiere a cuántas veces el valor de mercado cae fuera del intervalo de predicción, mientras que la respuesta media (valor esperado) cae dentro del intervalo de confianza para el mismo valor fijo de x.

Inferencias sobre los parámetros de la recta (95% de confianza):
Intervalo de confianza para b0: (49736.3561, 67855.4965)
Intervalo de confianza para b1: (0.5131, 0.5148)

Proporción de veces que el valor de mercado supera la incertidumbre de predicción:
Proporción de veces que el valor de mercado supera la incertidumbre de predicción: 0.0435

2 Ecuación de predicción del valor de mercado

b) Ecuación para predecir el valor de mercado del jugador a partir de varias características.

- i) Usando el método de mínimos cuadrados. Explica los indicadores obtenidos (como el coeficiente de determinación y la correlación) y proporciona una breve interpretación de los resultados.
- ii) Usando el método de descenso por gradiente. ¿Son los valores obtenidos iguales a los conseguidos mediante la resolución del sistema de ecuaciones normales? Muestra los resultados obtenidos junto con las últimas iteraciones del algoritmo. Indica los valores de los parámetros utilizados (como tasa de aprendizaje y número de iteraciones).
- iii) Da una interpretación del criterio de corte utilizado en el algoritmo del gradiente. Explica si presenta alguna falla. Si no es una buena condición de corte, ¿puedes sugerir un criterio alternativo más eficaz?

Mínimos cuadrados para regresión múltiple:
La regresión múltiple es una extensión de la regresión lineal simple en la que se utilizan varias variables predictoras $X_1, X_2, \ldots, X_p$ para predecir una variable respuesta Y. El modelo tiene la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Donde:

- Y es la variable dependiente (respuesta).
- $X_1, X_2, \ldots, X_p$ son las variables independientes (predictoras).
- β_0 es la ordenada al origen del modelo.
- $\beta_1, \beta_2, \ldots, \beta_p$ son los coeficientes de regresión asociados a cada predictor.
- ϵ es el error del modelo.

El método de mínimos cuadrados encuentra los valores de $\beta_0, \beta_1, \ldots, \beta_p$ que minimizan la suma de los errores cuadráticos entre los valores observados y los valores predichos por el modelo.

Indicadores clave obtenidos en la regresión múltiple:

Coeficiente de determinación R^2:
El coeficiente de determinación, conocido como R^2, mide qué tan bien el modelo ajusta los datos. Es la proporción de la varianza total en la variable

dependiente Y que es explicada por las variables independientes $X_1, X_2, \ldots, X_p$. Está dado por:

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

Donde:

- Y_i son los valores observados de la variable dependiente.
- $\hat{Y}_i$ son los valores predichos por el modelo.
- $\bar{Y}$ es la media de los valores observados.

Interpretación: Un valor de R^2 cercano a 1 indica que el modelo explica una gran proporción de la variabilidad de Y, mientras que un valor cercano a 0 indica que el modelo no explica bien la variabilidad en Y.

Coeficiente de correlación r:
El coeficiente de correlación es una medida de la fuerza y la dirección de la relación lineal entre dos variables. En regresión múltiple, la correlación no se puede medir de manera directa entre las variables predictoras y la respuesta, pero se puede medir la correlación entre el valor predicho y el valor observado de la variable dependiente.

Para una sola variable X y Y, el coeficiente de correlación se calcula como:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Interpretación: Un valor de r cercano a 1 indica una correlación lineal positiva fuerte, mientras que un valor cercano a -1 indica una correlación lineal negativa fuerte. Si $r = 0$, no hay correlación lineal entre las variables.

Valores obtenidos:

Coeficiente de determinación (($\mathbf{R^2}$)): 0.9885
Correlación(r): 0.9942

Coeficientes obtenidos

Métodos Mínimos Cuadrados:
Ordenada al origen (β_0): 0.0000113893667

potential (β_1): 0.00621768978
overall (β_2): 0.00948316611
release$_c lause_e ur$(β_3): 0.980412545

Descenso de gradiente:
Ordenada al origen (β_0): 0.0019961
potential (β_1): 0.00136138
overall (β_2): 0.00128694
release$_c lause_e ur$(β_3): 0.00028433

Podemos notar que son muy diferentes los valores obtenidos según el método utilizado, esto nos quiere decir que habría que ajustar la tasa de aprendizaje (disminuirla) así el número de iteraciones podría alcanzar valores más altos lo cual se traduce en mayor precisión, pero también implica mayor tiempo de ejecución. Esa sería la falla que encontramos en el criterio de corte, el criterio que planteamos es más eficaz.

3 Comportamiento del método de descenso por gradiente

c) Convergencia del método de descenso por gradiente. Explicar si el método siempre converge al mínimo de la función. En caso contrario, proporciona un contraejemplo para ilustrar este comportamiento.

No, no siempre converge al mínimo global de la función. ya que si usamos una función con al menos otro mínimo local además del mínimo global, puede ocurrir que el descenso por el gradiente nos lleve a un mínimo local y no al global.

Contraejemplo: Considerando la siguiente función no convexa: f(x)=x3x^3+2
Esta función tiene tanto un mínimo local como un mínimo global.

- El mínimo local ocurre cerca de x=0.5
- El mínimo global ocurre en x=2

Si empezamos el descenso por gradiente desde una condición inicial cerca de x=0.5, el método de descenso por gradiente se quedará atrapado en ese mínimo local en lugar de llegar al mínimo global en x=2.

Conclusión

Analizando y comparando los modelos de regresión lineal simple y múltilple, pudimos notar la similitud de ambos gráficos de dispersión: esto se debe a que la variable elegida en la regresión simple (La cláusula de recesión) tiene una fuerte correlación con el precio y las otras variables elegidas en la regresión múltiple no aportan información adicional significativa. Por lo tanto, podemos concluir que los dos modelos realizan predicciones similares. Esto explicaría por qué ambos gráficos se ven casi idénticos. También pudimos comparar los valores obtenidos entre los métodos de mínimos cuadrados y el descenso del gradiente analizando cuál criterio de corte es más eficaz y cuál es más eficiente, viendo las implicancias de las variaciones del valor de la tasa de aprendizaje.