Regresión Lineal

Aprenentatge Automàtic

APA/GEI/FIB/UPC - 2025/2026 1Q

/ Javier Béjar

- Hay dos tareas en aprendizaje supervisado: Regresión y Clasificación
- En regresión, la salida de cada ejemplo es un valor continuo, por lo que queremos un predictor definido como:

$$f(x) : \mathbb{R}^D \rightarrow \mathbb{R}$$

- Nada prohíbe que la salida sea multidimensional, pero solo trabajaremos con el caso unidimensional
- Generalmente supondremos que aprendemos de una muestra $X = \{x_1, \ldots, x_N\}$ de ejemplos iid con observaciones ruidosas $y_n = f(x_n) + \epsilon$ donde ϵ también es iid y modela la fuente de incertidumbre no observada

Regresión lineal

- Un **Modelo Lineal** será una función que es lineal en sus parámetros, no aparecen como exponentes, ni multiplicados/divididos por otros parámetros
- Eso no significa que sean funciones lineales, podemos combinar transformaciones no lineales de las características de entrada
- Por ejemplo, la regresión lineal se define como:

$$f(x) = w_0 + \sum_{d=1}^{D} w_d x_d$$

si consideramos un atributo adicional x_0 con valor constante 1, podemos expresar la función en notación vectorial como $f(x) = w^\top x$

- Podemos definir el problema de regresión lineal usando un enfoque probabilístico modelando el ruido como una distribución de probabilidad
- Consideraremos $\epsilon \sim \mathcal{N}(0, \sigma^2)$ iid con σ^2 conocida (homocedástica)
- Para el caso de regresión lineal nuestro modelo es:

$$p(y|x, w) = \mathcal{N}(y|w^\top x, \sigma^2)$$

- Observad que la media de la distribución depende de una función lineal de los datos
- Una predicción será una distribución de probabilidad donde obtendremos una estimación puntual como salida (la media)

- Podemos calcular las estimaciones de máxima verosimilitud de los parámetros usando la función de verosimilitud dado un conjunto de datos ($\mathcal{X}$,$\mathcal{Y}$)

$$p(\mathcal{Y}|\mathcal{X}, w) = \prod_{n=1}^{N} p(y_n|x_n, w) = \prod_{n=1}^{N} \mathcal{N}(y_n|w^\top x_n, \sigma^2)$$

- Minimizamos la log-verosimilitud negativa

$$-\log p(\mathcal{Y}|\mathcal{X}, w) = -\log \prod_{n=1}^{N} p(y_n|x_n, w) = -\sum_{n=1}^{N} \log p(y_n|x_n, w)$$

- Esto dará como resultado el método de mínimos cuadrados

- Dado que asumimos una distribución gaussiana, obtenemos[1]

$$\mathcal{L}(w) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - w^\top x_n)^2 + constante$$
$$= \frac{1}{2\sigma^2}(y - Xw)^\top(y - Xw) = \frac{1}{2\sigma^2}||y - Xw||_2^2$$

- Donde definimos X, la matriz de diseño, como $X = [x_1, x_2, \ldots, x_N]^\top \in \mathbb{R}^{N\times D}$, y y es el vector de respuesta
- Podemos definir la función de error para optimizar como

$$E(w) = \frac{1}{2}||y - Xw||_2^2$$

[1] Ver diapositiva 59 del primer tema del curso

- En este caso, la fórmula del error para encontrar un mínimo es una función cuadrática con respecto a los parámetros, por lo que podemos encontrar analíticamente un mínimo tomando derivadas e igualando a 0 (no hay que optimizar)
- Si el tamaño del problema es muy grande el coste computacional de obtener la solución exacta puede ser prohibitivo
- En ese caso se utilizan métodos de descenso de gradiente estocástico
- La solución no es exacta, pero es escalable

- Derivada respecto a los parámetros w

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w} &= \frac{\partial}{\partial w}\left(\frac{1}{2\sigma^2}(y - Xw)^\top (y - Xw)\right) \\ &= \frac{1}{2\sigma^2}\frac{\partial}{\partial w}\left(y^\top y - 2y^\top Xw + w^\top X^\top Xw)\right) \\ &= \frac{1}{\sigma^2}\left(-y^\top X + w^\top X^\top X\right)\end{aligned}$$

- Igualando la derivada a 0

$$w^\top X^\top X = y^\top X \Leftrightarrow w^\top = y^\top X(X^\top X)^{-1} \Leftrightarrow w = (X^\top X)^{-1}X^\top y$$

- El segundo paso asume que el rango de $X^\top X$ es D y, por lo tanto, es simétrica y definida positiva (y no es singular y tiene inversa)

- Resolver este problema requiere resolver un sistema de ecuaciones lineales (las ecuaciones normales) $Aw = b$ con $A = X^\top X$ y $b = X^\top y$
- Si es cierto que $X^\top X$ tiene rango D, entonces $(X^\top X)^{-1} X^\top$ es la matriz pseudoinversa X^+ o matriz de Moore-Penrose y $w = X^+ y$
- Problema: El coste de calcular la pseudoinversa es $O(N^3)$ e incluso si no es singular, puede causar problemas numéricos si está mal condicionada (cerca de ser singular)
- Para evitar estos problemas, estas ecuaciones se resuelven usando **Descomposición en Valores Singulares** (SVD)

- Cualquier matriz $A \in \mathbb{R}^{m \times n}$ con $m > n$ se puede expresar como:

$$A = U \Lambda V^\top$$

dónde

- $U \in \mathbb{R}^{m \times m}$ y $V \in \mathbb{R}^{n \times n}$ son matrices ortonormales con columnas de norma unitaria, por lo tanto $U^\top U = I$ y $V^\top V = I$ (U son los vectores propios $AA^\top$, V son los vectores propios de $A^\top A$)
- $\Lambda \in \mathbb{R}^{m \times n}$ es una matriz diagonal rectangular donde λ_i son los valores singulares (valores propios), si el rango de A es r entonces hay r valores singulares y el resto son 0

- Los parámetros w para mínimos cuadrados se pueden calcular usando SVD como:

$$w_i = V diag\left(\frac{1}{\lambda_i}\right)_+ U^\top y$$

donde U, V y Λ se calculan a partir de la SVD de X y

$$\left(\frac{1}{\lambda_i}\right)_+ = \max(0, \frac{1}{\lambda_i})$$

Este notebook muestra el cálculo de la regresión lineal usando la pseudoinversa y SVD

También podéis ver un video explicando el contenido del cuaderno

- La regresión lineal con los atributos de entrada solo puede ajustarse a líneas rectas
- Podemos calcular vectores de características a partir de los atributos de entrada aplicando diferentes funciones
- Podemos usar transformaciones no lineales para obtener funciones no lineales
- El modelo es lineal dado que todavía es lineal en los parámetros
- Generalizamos el modelo como:

$$y = f(x) = w^\top \phi(x) + \epsilon = \sum_{k=0}^{K} w_k \phi_k(x) + \epsilon$$

- ϕ se denominan funciones base

- Como hicimos antes, podemos calcular el estimador de máxima verosimilitud suponiendo ruido gaussiano

$$p(y|x, w) = \mathcal{N}(y|w^\top \phi(x), \sigma^2)$$

  donde $\phi : \mathbb{R}^D \to \mathbb{R}^K$ es una transformación no lineal de las entradas x
- Ahora los elementos de la matriz de diseño están compuestos por el resultado de esta transformación, que llamamos matriz de características $\Phi \in \mathbb{R}^{N \times K}$, donde $\Phi_{ij} = \phi_j(x_i)$
- El estimador de máxima verosimilitud corresponde a:

$$w_{ML} = (\Phi^\top \Phi)^{-1} \Phi^\top y$$

Regularización

- Como comentamos, el estimador de ML puede sobreajustar los datos
- Podemos usar la regularización para evitarlo
- En el caso de la regresión lineal, se puede controlar la magnitud de los pesos,
- Estos controlan cómo cambia la función entre dos ejemplos vecinos del conjunto de entrenamiento
- Queremos que la función cambie suavemente, para que no se desvíe mucho

- Para controlar la magnitud de los pesos podemos penalizar la función de error (error cuadrático)
- Una forma de controlar los pesos es agregar un término de penalización con su norma p

$$||w||_p^p = \left(\sum_{d=1}^{D} w_d^p \right)$$

- Las normas L_1 y L_2 se usan comúnmente en la práctica

- Agregar un término ponderado con la norma L_1 se denomina regresión LASSO y cuando $\lambda \to \infty$, más parámetros se reducen a 0 (solución dispersa (*sparse*))

$$E_\lambda(w) = \frac{1}{2}||y - \Phi w||^2 + \lambda \sum_{d=1}^{D} |w_d|$$

- Agregar un término ponderado con la norma L_2 se llama Regresión de Cresta (*Ridge Regression*) que tiende a reducir el valor de los parámetros (pero no suelen desaparecer)

$$E_\lambda(w) = \frac{1}{2}||y - \Phi w||^2 + \frac{\lambda}{2}||w||_2^2$$

- El parámetro λ debe ser mayor que 0 y se puede ajustar mediante validación cruzada

- Podemos combinar ambas regularizaciones en lo que se llama Elastic Net que usa una combinación convexa de ambas regularizaciones:

$$E_\lambda(w) = \frac{1}{2}||y - \Phi w||^2 + \lambda \left(\frac{1-\alpha}{2}||w||_2^2 + \alpha||w||_1^1 \right)$$

- Donde $\alpha \in [0, 1]$

Medición del ajuste

- Podemos usar el error cuadrático medio para comparar resultados de regresión

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(y_n - w^\top \phi(x_n))^2$$

- La raíz del error cuadrático medio da los resultados en las unidades de la variable de salida

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - w^\top \phi(x_n))^2}$$

- El coeficiente de determinación R^2 está normalizado

$$R^2 = 1 - \frac{MSE}{Var(y)}$$

- El Error Absoluto Medio no le da más importancia a los errores más grandes

$$MAE = \frac{1}{N}\sum_{n=1}^{N} |y_n - w^\top \phi(x_n)|$$

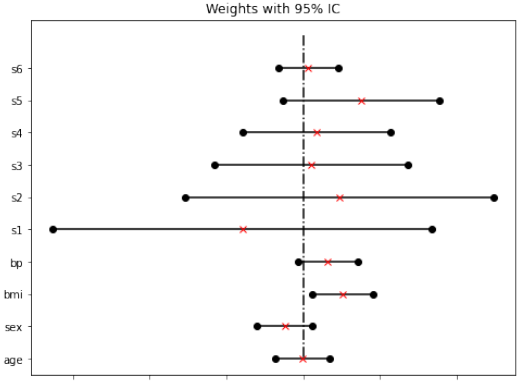
Esta función de error es más tolerante a los valores atípicos y se usa como objetivo de optimización en regresión robusta

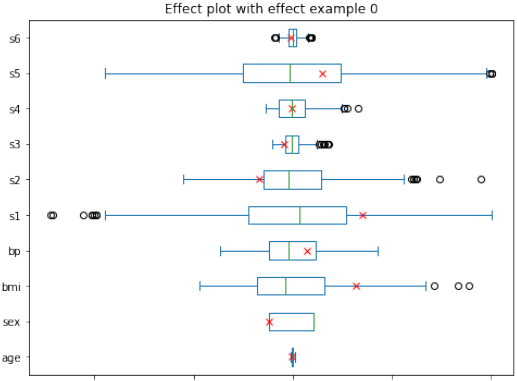
Interpretabilidad/Explicabilidad

- La regresión lineal es un modelo de caja blanca y proporciona una interpretación global del comportamiento del modelo
- Si los datos tienen muchas características, es mejor usar la regresión LASSO para tener un modelo disperso, más fácil de entender
- La importancia de las características se define como el peso dividido por el error estándar del peso, cuanto mayor sea el valor, mayor será la importancia
- El gráfico de efectos permite visualizar la distribución de la contribución de los valores de los datos a las predicciones

$$efecto_j^i = w_j x_j^i$$

- El efecto de las características en las predicciones individuales se puede explicar trazando los valores del ejemplo en el gráfico de efectos


Weights with 95% IC
s6
s5
s4
s3
s2
s1
bp
bmi
sex
age
−3000
−2000
−1000
0
1000
2000


Effect plot with effect example 0
s6
s5
s4
s3
s2
s1
bp
bmi
sex
age
−100
−50
0
50
100

- El poder predictivo del modelo es limitado dado que solo busca relaciones lineales
- Las no linealidades y las interacciones deben introducirse manualmente
- Las características también deben ser
  - Gausianas
  - Homocedásticas (varianza constante)
  - Independientes e idénticamente distribuidas (iid)
- Colinearidades: La importancia de las variables se distribuye a lo largo de todas las características correlacionadas

Este notebook aplica los modelos de regresión que hemos explicado al conjunto de datos Miles per Gallon