Tecnológico de Monterrey

Aprendizaje automático

Funciones de costo y optimización de parámetros

Dr. Omar Mendoza Montoya

14/08/2024

Problema de regresión

- Formalmente, un modelo de regresión es una función parametrizada que mapea un vector de entrada con un vector de salida.

$$\hat{Y} = f(X; \beta)$$

X = Vector de entrada con p variables

β = Vector de parámetros del modelo

$\hat{Y}$ = Vector de salida de m variables (numéricas reales)

Problema de regresión

- $\hat{Y}$ es el vector de salida del modelo, mientras que Y es el vector de valores verdaderos.

- Cada variable x_i en el vector X es llamada predictor o variable independiente, etc. Cada variable y_i en el vector Y es llamada variable dependiente o variable de criterio.

- A la diferencia entre Y y $\hat{Y}$ se le conoce como residuo r.

$$r = Y - \hat{Y}$$

- La pareja (X, Y) es una observación o muestra.

En regresión, se asume que las variables de salida tienen alguna relación o dependencia con las variables de entrada, por lo que el objetivo es determinar un modelo matemático que represente dicha relación.

Problemas en regresión

- Entrenamiento del modelo. Para una función regresora (función parametrizada), se deben encontrar los parámetros del modelo a través de un algoritmo de aprendizaje.
- Selección de modelo. Es necesario seleccionar un modelo de muchas posibilidades (selección de la función de regresión, sus predictores e hiperparámetros).
- Evaluación del modelo. Después de tener un modelo, el poder predictivo se debe estimar con un conjunto de datos independientes a los datos de entrenamiento.

Entrenamiento de modelos de regresión

Ejemplos de técnicas de aprendizaje para regresión

- Minimización de una función de costo o error
  - Mínimos cuadrados ordinarios (OLS)
  - Mínimos cuadrados ponderados
  - Mínimos cuadrados generalizados
- Regresión Bayesiana
- Regresión de cuantiles
- Regresión de componentes principales
- Regresión del ángulo mínimo

Independientemente de la técnica de aprendizaje, requerimos de un conjunto de datos de entrenamiento $D = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, donde cada pareja (X_i, Y_i) es una observación o muestra.

Estos datos se suelen representar por vectores o matrices según la cantidad de variables predictoras y variables dependientes.

Optimización por minimización del error

- Se define una función de costo, error o pérdida $C(D;\beta)$, la cual se calcula a partir de los datos con los que se entrena el modelo.

- La función de pérdida ayuda a comparar diferentes configuraciones del modelo $f(X;\beta)$ cambiando los valores del vector de parámetros β.

- La tarea es encontrar el valor de β^* que minimice la función de pérdida $C(D;\beta)$, es decir:

$$\beta^* = \underset{\beta}{\operatorname{argmin}}\, C(D;\beta)$$

Error cuadrático medio (MSE)

- El error cuadrático medio (mean squared error o MSE) es la función de error más utilizada para la optimización de parámetros en regresión.

Para una función $f(X_i; \beta)$, la función $MSE(D, \beta)$ se define por:

$$MSE(D, \beta) = \frac{1}{n}\sum_{i=1}^{n} \|Y_i - f(X_i; \beta)\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|r_i\|^2$$

Si Y_i solo tiene una variable y_i:

$$MSE(D, \beta) = \frac{1}{n}\sum_{i=1}^{n} \big(y_i - f(X_i; \beta)\big)^2 = \frac{1}{n}\sum_{i=1}^{n} (r_i)^2$$

Otras medidas de error

- Error absoluto medio (mean absolute error o MAE)

$$MAE(D,\beta) = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(X_i;\beta)| = \frac{1}{n}\sum_{i=1}^{n}|r_i|$$

- Función de pérdida Log-Cosh

$$L(D,\beta) = \frac{1}{n}\sum_{i=1}^{n}\log(\cosh(f(X_i;\beta) - y_i)) = \frac{1}{n}\sum_{i=1}^{n}\log(\cosh(-r_i))$$

Otras medidas de error

- Error absoluto medio suavizado (smooth mean absolute error, $SMAE$ o pérdida de Hubber)

$$SMAE(D,\beta) = \frac{1}{n}\sum_{i=1}^{n} L\big(y_i - f(X_i;\beta)\big) = \frac{1}{n}\sum_{i=1}^{n} L(r_i)$$

$$L(x) = \begin{cases} 0.5x^2 & \text{si } |x| \leq \delta \\ \delta|x| - 0.5\delta^2 & \text{en otro caso} \end{cases}$$

Para un valor δ pequeño.

¿Qué medida de error usar?

- Generalmente, MSE es una buena opción para entrenar el modelo, ya que tiene propiedades interesantes:
  - Si $f(X;\beta)$ es lineal, hay una solución analítica para MSE.
  - Si $f(X;\beta)$ no es lineal, la función de pérdida es convexa alrededor del óptimo, por lo que los algoritmos de optimización más típicos pueden ser utilizados.
- La solución por optimización de MSE se le conoce como solución por mínimos cuadrados ordinarios (OLS).
- Si es posible, se puede probar diferentes medidas de error para comprobar si hay diferencias en los modelos obtenidos.

¿Cómo optimizar una medida de error o costo?

- **Cálculo multivariado**

  Se calcula el gradiente ∇C de la función de costo C respecto a los parámetros del modelo β, se iguala a cero el gradiente, y se despejan los parámetros.

- **Métodos numéricos**

  Se utiliza un método iterativo con el cual se obtiene una aproximación de los parámetros β que minimizan la función de costo.

Optimización por descenso de gradiente

Descenso de gradiente

- Descenso de gradiente es un método de optimización numérica para funciones continuas y derivables.

- Consiste en iterar hasta converger utilizando la siguiente regla de actualización:

$$\beta \leftarrow \beta - \alpha \nabla C$$

donde α es una constante (usualmente pequeña) conocida como razón de aprendizaje.

Descenso de gradiente

1. Inicializa el vector β con valores pequeños aleatorios (ejemplo, entre -1 y 1).

2. Repetir hasta que $\|\nabla C\|$ sea pequeño (por debajo de un valor de umbral)

$$\beta \leftarrow \beta - \alpha \nabla C$$

La constante α puede ser un valor pequeño (0.0001) o se puede calcular para cada iteración.

Descenso de gradiente

- Si la función de costo es $MSE(D, \beta)$, note qué:

$$\nabla MSE(D, \beta) = \frac{1}{n}\sum_{i=1}^{n} \nabla (r_i)^2 = \frac{2}{n}\sum_{i=1}^{n} (r_i)\nabla r_i$$

- Para la función $MAE(D, \beta)$, no se puede aplicar el método de descenso de gradiente, ya que no es derivable en todo punto.

- En general, se debe tener cuidado con la selección de la función de costo de tal manera que ésta se pueda optimizar con un método basado en el gradiente.

Gradiente para regresión lineal múltiple

- Para el caso de regresión lineal múltiple:

$$r_i = y_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \qquad \nabla r_i = -\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = -X_i$$

$$\nabla MSE(D, \beta) = -\frac{2}{n}\sum_{i=1}^{n} r_i X_i$$

- Regla de aprendizaje:

$$\beta \leftarrow \beta + \alpha\left(\frac{2}{n}\sum_{i=1}^{n} r_i X_i\right)$$

Bibliografía

- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2023). *An introduction to statistical learning: with applications in Python* (2da ed.). Springer.
  - Capítulo 3

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2da ed.). Springer.
  - Capítulo 3

- Russell, S. J. & Norvig, P. (2021). *Artificial intelligence: A modern approach (global edition)* (4ta ed.). Person.
  - Capítulo 19