5. Optimización Bayesiana

La optimización Bayesiana a diferencia de otros métodos como grid search o random search, es una forma de buscar los mejores hiperparámetros de un modelo mediante el uso de 2 herramientas, que, tras varias iteraciones, logra llegar a estos hiperparámetros deseados. Los problemas que trata este método son del tipo:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{A}} f(\mathbf{x})$$

donde:

- f es la función objetivo
- $\mathcal{A}$ es el espacio de búsqueda o conjunto de parámetros.

El objetivo es encontrar el conjunto de parámetros $\mathbf{x}^*$ que maximiza (o minimiza) $f(\mathbf{x})$.

El proceso es secuencial y se basa en el Teorema de Bayes para guiar la búsqueda, equilibrano entre explotación (μ alta) y exploración (σ alta).

5.1. Surrogate Modelo

Generalmente, se utiliza un proceso gaussiano, este es un modelo que se ajusta a los puntos observados o evaluados para aproximar la función objetivo desconocida, sin embargo, no es un simple ajuste, ya que lo que proporciona es una distribución de probabilidad alrededor de este ajuste, es decir, calcula una media y una varianza de tal forma que se tienen infinitas funciones que se ajustan a los puntos conocidos que viven dentro de la distribución calculada.

Se utiliza como una distribución previa (prior) sobre la función f. Cada nueva evaluación de f (el "dato") se usa para actualizar la previa y obtener una distribución posterior (posterior) más precisa, usando la regla de Bayes.

En resumen, un GP define una distribución sobre funciones:

$$f \sim \mathcal{GP}(m(x), k(x, x'))$$

Donde,

- $m(x)$ es la función media
- $k(x, x')$ es la función de covarianza o kernel

Entonces, dado un conjunto de evaluaciones/observaciones:

$$\mathcal{D}_n = \{(x_i, f(x_i))\}_{i=1}^n$$

El GP da una distribución posterior:

$$f(x)|\mathcal{D}_n \sim \mathcal{N}(\mu(x), \sigma^2(x))$$

Donde,

- $\mu(x)$ altos indican regiones a explotar. Seguridad de obtener buenos resultados
- $\sigma^2(x))$ altos indican regiones a explorar. Regiones desconocidas, posibilidad de ser una buena o mala region.

5.2. Función de adquisición

Es una función más fácil de evaluar que la función objetivo, utiliza μ y σ^2 del surrogate model para calcular el siguiente punto a evaluar en la función objetivo. Esta función hace un balance entre exploración y explotación. Generalmente se utiliza el Expected Improvement (EI). Esta función mide el valor esperado de mejora sobre el mejor punto observado.

$$EI(x) = \mathbb{E}[\max(f^*(x) - f(x), 0)]$$

5.3. Implementación con Optuna

En la implementación, se suele utilizar la librería Optuna, esta librería trabaja con el surrogate model: TPE (Tree-Structured Parzen Estimator) Este es un modelo no parámetrico basado en los estimadores de Parzen. En este caso, en vez de modelar:

$$p(f|x)$$

Se modela:

$$p(x|f)$$

Y luego, utilizando el teorema de Bayes se selecciona el nuevo punto a evaluar. Optuna separa las evaluaciones en dos grupos:

- $l(x)$: mejores evaluaciones
- $g(x)$: peores evaluaciones

Luego aprende dos estimadores de densidad:

$$l(x) = p(x|f \leq f^*)$$

$$g(x) = p(x|f > f^*)$$

Donde f^* es un umnbral. Estas densidades se modelan a partir de estimadores de Parzen.

Para seleccionar el siguiente punto, Optuna maximiza el siguiente cociente:

$$\frac{l(x)}{g(x)}$$

Es decir, se buscan los valores de hiperparámetros que son típicos en configuraciones buenas y atípicos en configuraciones malas. En este caso, la función de adquisición no está definida explicitamente, sino que se implementa a través del crietio antes mencionado, es decir:

$$x_{n+1} = \mathbf{arg}\max_x \frac{l(x)}{g(x)}$$