Estudio del método híbrido de gradiente estocástico y ecuación de Langenvin

Andaur, Victoria[1]; **Vera, Matias**[2]

Profesor: Joaquin Fontbona T. ; Auxiliares: Pablo Araya; Bruno Hernandez

Departamento de Ingeniería matemática, FCFM, Universidad de Chile

20 de diciembre de 2021

Abstract: El proyecto busca estudiar el impacto del método híbrido de en el paper "Bayesian Learning via Stochastic Gradient Langevin Dynamics"[1] , para ello se busca replicar los experimentos realizados en [1] para distintos set de data de "juguete", especialmente aquellos estudiados en el laboratorio 4.1 del curso de simulación estocástica MA4402 [2].

Palabras clave: Aprendizaje Bayesiano, Distribución posterior, Gradiente descendiente estocástico, Ecuación de Langevin.

1. Introducción

Como sociedad, en casi todas las áreas de la vida, nos interesa realizar predicciones con datos, pero en los últimos años ha habido un incremento masivo en la cantidad de datos disponible lo que ha conllevado que se manejen constantemente conjuntos de datos que superan el millón de casos y es precisamente en estos conjuntos en los que los métodos de Cadenas de Markov de Monte Carlo (MCMC) fallan, ya que para implementarse requieren actuar sobre la totaliad de los datos. Por otro lado, el método de gradiente descendiente estocástico (SGD) permite trabajar con más información de manera rápida, pero no permite cuantificar la incerteza de los datos. Es así que [1] propone un modelo híbrido, en donde se pueda calcular la incerteza de los datos a un bajo costo computacional.

Se busca optimizar θ, vector de parámetros, a través de un conjunto de datos $X = \{x_{t1}, \cdots, x_{tn}\}$. El método de SGD propone:

$$\Delta\theta_t = \frac{\epsilon_t}{2}\Big(\nabla log\ p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla log\ p(x_{ti}|\theta_t)\Big) \quad (1)$$

donde ϵ_t suman infinito, pero sus cuadrados tienen suma finita.

El paper [1] propone que podemos interpretar la optimización de la trayectoria del SGD como una cadena de Markov con una distribución del equilibrio sobre la distribución posterior sobre θ. Esto quiere decir que podemos entrenar el modelo usando SGD regular y agregar un ruido normal en cada paso. Después hacemos que la taza de aprendizaje y el ruido inducido tiendan a 0 cuando el tiempo t tiende a ∞. De manera intuitiva, esto funciona ya que permite que la optimización encuentre un mínimo local, pero nunca va a converger debido al ruido inducido, pero también previene que se vaya del mínimo local gracias al decaimiento de la tasa de aprendizaje. El nuevo método es una combinación del método presentado en 1, con la ecuación de Lagensive 2:

$$\Delta\theta_t = \frac{\epsilon}{2}\Big(\nabla log\ p(\theta_t) + \sum_{i=1}^{N} \nabla log\ p(x_i|\theta_t)\Big) + \eta_t \quad (2)$$

donde $\eta_t \sim \mathcal{N}(0, \epsilon)$ Dando como resultado, el método SGLD:

$$\Delta\theta_t = \frac{\epsilon_t}{2}\Big(\nabla log\ p(\theta_t) + \frac{N}{n}\sum_{i=1}^{N} \nabla log\ p(x_t i|\theta_t)\Big) + \eta_t \quad (3)$$

donde $\eta_t \sim \mathcal{N}(0, \epsilon_t)$ y $\epsilon_t = a(b+t)^{-\gamma}$

2. Metodología

Se replicaron los experimentos presentados en [1] con los presentados en [2]

Referencias

[1] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 681–688.

[2] Laboratorio 4.1 curso MA4402-2 fcfm.