Масштабируемая Аппроксимация Лапласа для нейронных сетей

Колесов Александр
Московский Физико-Технический Институт

Байесовское мультимоделирование
29.09.2021

Ряд Тейлора

Теорема Тейлора(одномерный случай)

Пусть задана функция $f \in C^{n+1}([a,b])$, тогда для любой произвольной точки $\forall x_0 \in [a,b]$ выполняется следующее разложение с некоторым остаточным членом

$$f(x) = f(x_0) + ... + \frac{f^n(x_0)}{n!}(x - x_0)^n + R_n$$

Теорема Тейлора(многомерный случай)

В случае когда функция представляет собой функцию многих переменных , то разложение до второго порядка представляет собой следующий вид, при сохранении ограничений предыдущей теоремы

$$f(x) = f(x_0) + \nabla f(x_0)^T + \frac{1}{2}(x - x_0)^T H(x - x_0)$$

Аппроксимация Постериорного распределения

- $p(\phi|D)$ - апостериорное рапсределение на веса модели при наличии данных D
- Poor bayesian inference - можно получить моду постериорного распределения
- Можно воспользоваться разложением Тейлора

$$\log p(\phi^*|D) + \nabla \log p(\phi^*|D)^T(\phi - \phi^*) - \frac{1}{2}(\phi - \phi^*)^T H(\phi - \phi^*)$$

- Экспоненциируем выражение

$$\log p(\phi|D) \sim \log p(\phi^*|D) - \frac{1}{2}(\phi - \phi^*)^T H(\phi - \phi^*)$$

- Получаем , что $\phi \sim \mathcal{N}(\phi^*, H^{-1})$

Матрица Фишера

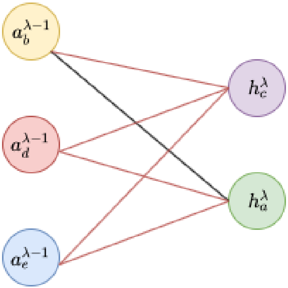
- Обращать Гессиан тяжело вычислительно
- матрица Фишера $F = \mathbb{E}(\nabla_\phi \log p(y|x) \nabla_\phi \log p(y|x)^T)$
- $H \sim diag(F) = diag(\mathbb{E}[(\nabla_\phi \log p(y|x))^2])$

Такой подход позволяет нам для каждого слоя λ получать веса, как семпл из нормального распределения, с приведенной матрицей Фишера

$$vec(W_\lambda) \sim \mathcal{N}(vec(W_\lambda^*), diag(F_\lambda)^{-1})$$

Однако, такой подход не учитывает потенциальной значимой ковариации между весами даже в одном слое

Немного нотации



- $a^{\lambda-1}$ - прошлые активации
- $h^{\lambda} = W^{\lambda} a^{\lambda-1}$
- $f(h^{\lambda}) = a^{\lambda}$ -новые активации
- $w_{a,b}^{\lambda}$ - вес

Вычисление Гессиана

Прежде всего напомним определение Гессиана

$$H_{ij} = \frac{\partial^2}{\partial \phi_i \partial \phi_j} \mathbb{E}(\log p_\phi(y|x))$$

Научимся вычислять первую производную при помощи chain rule:

$$\frac{\partial E}{\partial w^\lambda_{a,b}} = \sum_i \frac{\partial E}{\partial h^\lambda_a} \frac{\partial h^\lambda_a}{\partial w^\lambda_{a,b}} = \frac{\partial E}{\partial h^\lambda_a} a^{\lambda-1}_b$$

Теперь попрубуем посчитать вторую производную

$$[H_\lambda]_{(a,b),(c,d)} = \frac{\partial^2 E}{\partial w^\lambda_{a,b} \partial w^\lambda_{c,d}} = a^{\lambda-1}_b a^{\lambda-1}_d \frac{\partial^2 E}{\partial h^\lambda_a \partial h^\lambda_c}$$

Факторизация Кронекера

Переходя от элементов к векторным величина получаем следующие выражения, для ковариации входных активаций

$$\mathcal{G}_\lambda = a_{\lambda-1} a_{\lambda-1}^T$$

И для Гессиана по пре-активациям слоя

$$\mathcal{H}_\lambda = \frac{\partial E}{\partial h^\lambda \partial h^\lambda}$$

Поскольку, для каждых входных нейронов нам надо рассмотреть все возможные выходные нейроны, то получается Фактор Кронекера

$$H_\lambda = \mathcal{G}_\lambda \otimes \mathcal{H}_\lambda$$

Факторизация Кронекера

- Делаем факторизацию по слоям
- Гессиан представляет из себя блочно-диагонанльную матрицу
- матрицы меньших размеров теперь D^2 вместо D^4
- Свойство Гессиана

$$H_\lambda^{-1} = \mathcal{G}_\lambda^{-1} \otimes \mathcal{H}_\lambda^{-1}$$

Фактор Кронекера для Аппроксимации Лапласа

Возвращаемся к проблеме вычисления обратного гессиана для семплирования из нормального распределения

- (Gupta et al,2019) факторизация матрицы ковариации по строкам и столбцам
- Рассмотрим факторизацию Кронекера в нашем случае

$$vec(W_\lambda) = \mathcal{N}(vec W^*_\lambda, \mathcal{G}_\lambda^{-1} \otimes \mathcal{H}_\lambda^{-1})$$

- Предполагается независимость $\mathcal{G}, \mathcal{H}$

$$\mathbb{E}(H_\lambda) \sim \mathbb{E}(\mathcal{H}_\lambda) * \mathbb{E}(\mathcal{G}_\lambda)$$