Chú thích cho phần "Exact Greedy Algorithm for Split Finding" trong bài báo "XGBoost: A Scalable Tree Boosting System"

Để huấn luyện mô hình XGBoost, chúng ta cần quan tâm đến hàm mất mát:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, F_i^{(t-1)} + w_{x_i}^{(t)}\right) + \gamma T^{(t)} + \frac{1}{2}\lambda \sum_{j=1}^{T^{(t)}} \left(w_j^{(t)}\right)^2 \tag{1}$$

Trong đó, t biểu thị cây quyết định thứ t, n là số lượng mẫu huấn luyện, và l là hàm mất mát, có thể là *mean square error* cho hồi quy hay *binary cross entropy* cho phân loại nhị phân. Tham số w_j được sử dụng để biểu diễn giá trị tại nút lá thứ j trong cây có T nút lá.
Nhắc lại khai triển Taylor:

$$f(x+a, y+b) = f(x,y) + \left(\frac{\partial f}{\partial x}a + \frac{\partial f}{\partial y}b\right) + \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^2}a^2 + \frac{\partial^2 f}{\partial y^2}b^2 + 2\frac{\partial^2 f}{\partial x \partial y}ab\right) + \dots \tag{2}$$

Áp dụng khai triển Taylor bậc hai cho hàm mất mát:

$$l\left(y_i, F_i^{(t-1)} + w_{x_i}^{(t)}\right) = l\left(y_i, F_i^{(t-1)}\right) + \frac{\partial l}{\partial F_i^{(t-1)}} w_{x_i}^{(t)} + \frac{1}{2}\frac{\partial^2 l}{\partial \left(F_i^{(t-1)}\right)^2}\left(w_{x_i}^{(t)}\right)^2 + \dots \tag{3}$$

Để tiện cho việc tính toán sau này, đặt ${g_i}^{(t)}$ và ${h_i}^{(t)}$ như sau:

$${g_i}^{(t)} = \frac{\partial l}{\partial F_i^{(t-1)}} \qquad {h_i}^{(t)} = \frac{\partial^2 l}{\partial \left(F_i^{(t-1)}\right)^2} \tag{4}$$

Như vậy, (1) trở thành:

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}\left[l\left(y_i, F_i^{(t-1)}\right) + {g_i}^{(t)} w_j^{(t)} + \frac{1}{2}{h_i}^{(t)}\left(w_j^{(t)}\right)^2\right] + \gamma T^{(t)} + \frac{1}{2}\lambda \sum_{j=1}^{T^{(t)}}\left(w_j^{(t)}\right)^2 \tag{5}$$

$$\simeq \sum_{j=1}^{T}\left[\left(\sum_{i \in I_j} {g_i}^{(t)}\right)\left(w_j^{(t)}\right) + \frac{1}{2}\left(\sum_{i \in I_j} {h_i}^{(t)} + \lambda\right)\left(w_j^{(t)}\right)^2\right] + \gamma T^{(t)} \tag{6}$$

Xét một cây quyết định có cấu trúc cố định tại thời điểm t, chúng ta có thể xác định ${w_j^{(t)}}^*$ sao cho giá trị $\mathcal{L}^{(t)}$ là tối thiểu:

$${w_j^{(t)}}^* = -\frac{\sum_{i \in {I_j}^{(t)}} {g_i}^{(t)}}{\sum_{i \in {I_j}^{(t)}} {h_i}^{(t)} + \lambda} \tag{7}$$

Và thay ${w_j^{(t)}}^*$ vào (6), ta được:

$$\tilde{L}^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\left(\frac{\sum_{i \in {I_j}^{(t)}} {g_i}^{(t)}}{\sum_{i \in I_j}^{(t)} {h_i}^{(t)} + \lambda}\right)^2 + \gamma T^{(t)} \tag{8}$$

Tuy nhiên, việc sử dụng (8) làm tiêu chí để chọn cây quyết định tối ưu tại thời điểm t là không khả thi, vì trước khi tách nút, chúng ta không thể biết trước số lượng nút của cây (giá trị $T^{(t)}$). Mục đích là tìm ra cây tối ưu cho thời điểm t. Có thể tạo ra tất cả các cây quyết định có thể và chọn cây có giá trị $\tilde{L}^{(t)}$ thấp nhất, nhưng phương pháp này quá phức tạp và bất khả thi. Do đó, chúng ta có thể áp dụng một phương pháp xấp xỉ, sử dụng một thuật toán tham lam bằng cách biến đổi (8) thành tiêu chí (tiêu chí này có thể được coi là tương đương với chỉ số Gini hoặc Entropy được sử dụng trong phương pháp CART) để phân tách nút khi xây dựng cây quyết định.

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{9}$$

Trong đó, I là tập hợp các mẫu tại nút cần được tách thành hai, I_L là tập hợp các mẫu tại nút con bên trái sau tách, và I_R là tập hợp các mẫu tại nút con bên phải sau tách. Chúng ta sẽ bỏ qua T trong (8) vì tổng số nút lá của cây quyết định không thể xác định trước khi cây được hoàn thành.

Thuật toán 1: Exact Greedy Algorithm for Split Finding

Đầu vào: I, tập thể hiện của nút hiện tại
Đầu vào: d, chiều của đặc trưng
điểm_cao_nhất $\leftarrow 0$
$G \leftarrow \sum_{i\in I} g_i$
$H \leftarrow \sum_{i\in I} h_i$
for $k = 1$ đến d **do**
    $G_L \leftarrow 0$
    $H_L \leftarrow 0$
    for j trong I đã sắp xếp theo x_{jk} **do**
        $G_L \leftarrow G_L + g_j$
        $H_L \leftarrow H_L + h_j$
        $G_R \leftarrow G - G_L$
        $H_R \leftarrow H - H_L$
        điểm_mới $\leftarrow \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda}$
        điểm_cao_nhất $\leftarrow \max($điểm_cao_nhất, điểm_mới$)$
Đầu ra: Chia điểm cao nhất điểm_cao_nhất

Để cài đặt thuật toán, ta cần tính toán chi tiết các giá trị ${g_i}^{(t)}$ và ${h_i}^{(t)}$. Trong bối cảnh bài toán hồi quy, các giá trị này được định tính như sau:

$$F_i^{(t)} = {\hat{y}_i}^{(t)} \tag{10}$$

$$l\left(y_i, F_i^{(t)}\right) = \left(y_i - {\hat{y}_i}^{(t)}\right) \tag{11}$$

$$g_i{}^{(t)} = \frac{\partial l}{\partial F_i^{(t-1)}} = \frac{\partial}{\partial \hat{y}_i{}^{(t-1)}} \left[\frac{1}{2} \left(y_i - \hat{y}_i{}^{(t-1)} \right)^2 \right] \tag{12}$$

$$g_i{}^{(t)} = - \left(y_i - \hat{y}_i{}^{(t-1)} \right) \tag{13}$$

$$h_i{}^{(t)} = \frac{\partial^2 l}{\partial \left(F_i^{(t-1)} \right)^2} = \frac{\partial^2}{\partial \left(\hat{y}_i{}^{(t-1)} \right)^2} \left[\frac{1}{2} \left(y_i - \hat{y}_i{}^{(t-1)} \right)^2 \right] \tag{14}$$

$$h_i{}^{(t)} = 1 \tag{15}$$

Đối với boài toán phân loại nhị phân, chúng ta tìm cách đưa bài toán phân loại nhị phân về bài toán hồi quy:

$$F_i^{(t)} = \log \left(\frac{\hat{y}_i{}^{(t)}}{1 - \hat{y}_i{}^{(t)}} \right) \tag{16}$$

$$\hat{y}_i{}^{(t)} = \frac{1}{1 + \exp\left(-F_i^{(t)} \right)} \tag{17}$$

$$l \left(y_i, F_i^{(t)} \right) = -y_i F_i^{(t)} + \log \left(1 + \exp \left(F_i^{(t)} \right) \right) \tag{18}$$

$$g_i{}^{(t)} = \frac{\partial l}{\partial F_i^{(t-1)}} = \frac{\partial}{\partial F_i^{(t-1)}} \left[-y_i F_i^{(t-1)} + \log \left(1 + \exp \left(F_i^{(t-1)} \right) \right) \right] \tag{19}$$

$$g_i{}^{(t)} = -y_i + \frac{\exp \left(F_i^{(t-1)} \right)}{1 + \exp \left(F_i^{(t-1)} \right)} = -y_i + \frac{1}{1 + \exp \left(-F_i^{(t-1)} \right)} \tag{20}$$

$$g_i{}^{(t)} = - \left[y_i - \frac{1}{1 + \exp \left(-F_i^{(t-1)} \right)} \right] \tag{21}$$

$$g_i{}^{(t)} = - \left(y_i - \hat{y}_i{}^{(t-1)} \right) \tag{22}$$

$$h_i{}^{(t)} = \frac{\partial^2 l}{\partial \left(F_i^{(t-1)} \right)^2} = \frac{\partial}{\partial F_i^{(t-1)}} \left[-y_i + \frac{1}{1 + \exp \left(-F_i^{(t-1)} \right)} \right] \tag{23}$$

$$h_i{}^{(t)} = \frac{\exp \left(-F_i^{(t-1)} \right)}{\left[1 + \exp \left(-F_i^{(t-1)} \right) \right]^2} = \frac{\exp \left(-F_i^{(t-1)} \right)}{1 + \exp \left(-F_i^{(t-1)} \right)} \cdot \frac{1}{1 + \exp \left(-F_i^{(t-1)} \right)} \tag{24}$$

$$h_i{}^{(t)} = \left[\frac{1 + \exp \left(-F_i^{(t-1)} \right)}{1 + \exp \left(-F_i^{(t-1)} \right)} - \frac{1}{1 + \exp \left(-F_i^{(t-1)} \right)} \right] \cdot \frac{1}{1 + \exp \left(-F_i^{(t-1)} \right)} \tag{25}$$

$$h_i{}^{(t)} = \hat{y}_i^{(t-1)} \left(1 - \hat{y}_i^{(t-1)} \right) \tag{26}$$