$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$ 损失函数+正则化项

对于 $\sum_{i=1}^{n} l(y_i, \hat{y}_i)$，在第 t 次迭代中，$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$，$f_t(x_i)$ 是一个新的树的叶子节点上的权重

$$= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))$$

由 Taylor 展开：设 $l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) = F(\hat{y}_i^{(t-1)} + f_t(x_i))$

$$\approx F(\hat{y}_i^{(t-1)}) + F'(\hat{y}_i^{(t-1)}) \cdot f_t(x_i) + \frac{1}{2} F''(\hat{y}_i^{(t-1)}) f_t^2(x_i)$$

so $l(y_i, \hat{y}_i^{(t)}) = l(y_i, \hat{y}_i^{(t)}) + \frac{\partial l(y_i, \hat{y}_i^{(t)})}{\partial \hat{y}_i^{(t-1)}} \cdot f_t(x_i) + \frac{1}{2} \frac{\partial^2 l(y_i, \hat{y}_i^{(t)})}{\partial \hat{y}_i^{(t-1)^2}} \cdot f_t^2(x_i)$ (☆)

对于 $\sum_{k=1}^{K} \Omega(f_k) = \sum_{k=1}^{K-1} \Omega(f_k) + \Omega(f_K)$ (☆)

将两个☆代入原本的 Obj 后：

$$Obj(\theta) = \sum_{i=1}^{n} \underbrace{l(y_i, \hat{y}_i^{(t)})}_{\text{常数项}} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) + \underbrace{\sum_{k=1}^{t-1} \Omega(f_k)}_{\text{常数项}} + \Omega(f_t)$$

则我们得到了第 t 步带零化的目标函数：

$$Obj^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

它使得 XGBoost 可以支持任何可二次求导的损失函数，即不仅仅是均方误差，在这个式子中，g_i, h_i 只与传统的损失函数相关，核心部分是我们希望求解的树 f_t。

对于一颗给定的树结构，假设我们将样本分到了 T 个叶子节点上，每个叶子节点 j 包含了样本集合 I_j。当一个样本 x_i 输入到第 K 棵树时，它会沿着树的路径最终到达一个叶子节点，我们将这个叶子节点表示为 $q(x_i)$。这个叶子节点有一个预测的"分数"或"预测值"，我们用 $w_{q(x_i)}$ 来表示。因此，第 K 棵树对样本 x_i 的最终预测结果 $f_k(x_i)$，就等于样本 x_i 所落入的那个叶子节点的分数 $w_{q(x_i)}$，即 $f_k(x_i) = w_{q(x_i)}$。这是对于每一个样本而言的叶子节点权重，在同一个叶子节点上的所有样本对应的叶子权重是相同的

基于以上这些信息，我们定义第k棵树 f_k 的复杂度 $\Omega(f_k)$

$\Omega(f_k) := \gamma T + \frac{1}{2}\alpha\sum_{j=1}^{T}|w_j| + \frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$，$w_j$ 为第 j 个叶子节点上的预测分数，γ 为控制叶子节点数量的惩罚项系数，λ 为控制叶子节点权重的 L_2 正则化系数。

α ——— L_1 ———

if $q(x_i) \in I_j$, then $w_{q(x_i)} = w_j := f_k(x_i)$.

当 $\alpha = \lambda = 0$ 时，则为普通的GBDT

下面我们只考虑 L_2 正则化：

$$\text{then } Obj^{(t)} = \sum_{i=1}^{m}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$$

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

$$\frac{\partial Obj^{(t)}}{\partial w_j} = \sum_{i\in I_j}g_i + w_j\sum_{i\in I_j}h_i + \lambda w_j = 0 \qquad \therefore w_j^* = \frac{-\sum_{i\in I_j}g_i}{\lambda + \sum_{i\in I_j}h_i}$$

这样我们就得到了每个叶子节点 j，其最优的预测分数 w_j^*

将 w_j^* 代回 Obj，则我们有 $Obj^{(t)}(\text{最优}) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j}g_i)^2}{\sum_{i\in I_j}h_i + \lambda} + \gamma T$

这时候，样本已经被归结到了每个叶子中

那么，目标函数就变为基于树的结构来进行计算。

这样，我们就建立了树的结构（叶子）和模型效果的直接关系。

我们的目标函数又叫（结构分数）(Structure Score)