TG-FOBOS-RDA-FTRL 算法

1. 简单截断法与 TG 算法

在基于 SGD 的逻辑回归中，第 t+1 次样本的第 i 个特征的权重更新公式为：

$$w_i^{t+1} = w_i^t - \eta_i^t g_i^t$$

其中，η_i^t 是第 t 次第 i 个特征学习率，g_i^t 是第 t 次第 i 个特征梯度。

为了防止过拟合，除了采用 L1 正则化和 L2 正则化外，还可以采用简单截断法。在简单截断法中，以 k 为窗口，当 t/k 不为整数时，采用上式更新权重，当 t/k 为整数时，采用如下公式更新权重：

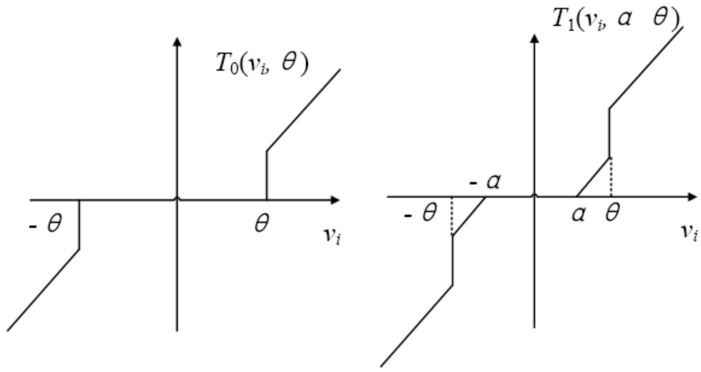
$$w_i^{t+1} = T_0(v_i, \theta)$$
$$T_0(v_i, \theta) = \begin{cases} 0 & |v_i| \le \theta \\ v_i & otherwise \end{cases}$$

上式中，θ>0，$v_i = w_i^t - \eta_i^t g_i^t$。

但是由于简单截断法太过于粗暴，因此 TG(截断梯度法，Truncated Gradient)在此基础上做了改进，显得不那么粗暴。同样以 k 为窗口，当 t/k 不为整数时，更新方式相同，当 t/k 为整数时，采用如下公式更新权重：

$$w_i^{t+1} = T_1(v_i, \alpha, \theta)$$
$$T_1(v_i, \alpha, \theta) = \begin{cases} \max(0, v_i - \alpha) & 0 \le v_i < \theta \\ \max(0, v_i + \alpha) & -\theta \le v_i < 0 \\ v_i & otherwise \end{cases}$$

上式中，θ>0，$v_i = w_i^t - \eta_i^t g_i^t$，0<$\alpha$<$\theta$。通过上面的对比，我们可以发现简单截断法与 TG 算法的不同，对比如下图所示：



通过上图可以发现：在 TG 算法中，当 α=θ 时，TG 算法就是简单截断法。

2. FOBOS 算法

前向后向切分(FOBOS, Forward-Backward Splitting)是由 John Duchi 和 Yoram Singer 提出的。从全称上来看，该方法应该叫 FOBAS，但是由于一开始作者管这种方法叫

FOLOS(Forward Looking Subgradients)，为了减少读者的困扰，作者干脆只修改一个字母，叫 FOBOS。在 FOBOS 中，权重向量的更新方式分为两个步骤：

$$\boldsymbol{W}^{t+\frac{1}{2}} = \boldsymbol{W}^{t} - \boldsymbol{\eta}^{t}\boldsymbol{g}^{t}$$

$$\boldsymbol{W}^{t+1} = \underset{\boldsymbol{W}}{\operatorname{argmin}}\left\{\frac{1}{2}\left\|\boldsymbol{W} - \boldsymbol{W}^{t+\frac{1}{2}}\right\|^{2} + \Psi(\boldsymbol{W})\right\}$$

上式中，$\Psi(\boldsymbol{W})$为 L1 正则化与 L2 正则化。将其拆分为第 i 个特征维度上，即：

$$w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left(\frac{1}{2}(w_i - v_i)^2 + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2\right)$$

上式中 $v_i = w_i^t - \eta_i^t g_i^t$，$\lambda_1$ 和 λ_2 为 L1 正则化与 L2 正则化的系数。假设 w_i^* 是上式的最优解，

所以 $w_i^* v_i \geq 0$，这是因为：

反证法：

假设：$w_i^* v_i < 0$，那么有：

$$\frac{1}{2}v_i^2 < \frac{1}{2}v_i^2 - w_i^* v_i + \frac{1}{2}(w_i^*)^2 < \frac{1}{2}(w_i^* - v_i)^2 + \lambda_1|w_i^*| + \frac{1}{2}\lambda_2(w_i^*)^2$$

这与 $\underset{w_i}{\operatorname{argmin}}\left(\frac{1}{2}(w_i - v_i)^2 + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2\right)$ 最优解相矛盾，所以假设不成立。

既然 $w_i v_i \geq 0$，那么分两种情况 $v_i \geqslant 0$ 和 $v_i < 0$ 来讨论：

(1)当 $v_i \geqslant 0$ 时：

由于 $w_i v_i \geq 0$，所以 $w_i \geqslant 0$，相对于在 $\underset{w_i}{\operatorname{argmin}}\left(\frac{1}{2}(w_i - v_i)^2 + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2\right)$ 中引入了不等式的约束条件 $-w_i \leqslant 0$，引入拉格朗日乘子 $\beta \geqslant 0$，由 KKT 条件(参考备注)有：

$$\begin{cases} \dfrac{\partial}{\partial w_i}\left(\dfrac{1}{2}(w_i - v_i)^2 + \lambda_1|w_i| + \dfrac{1}{2}\lambda_2 w_i^2 - \beta w_i\right) = 0 \\ \beta w_i = 0 \end{cases}$$

所以：$w_i = \dfrac{1}{1+\lambda_2}(v_i - \lambda_1 + \beta)$，分两种情况：

a.当 $w_i > 0$ 时：

由于 $\beta w_i = 0$，所以 $\beta = 0$，$w_i = \dfrac{1}{1+\lambda_2}(v_i - \lambda_1) > 0$

b.当 $w_i = 0$ 时：

由于 $\beta \geqslant 0$，所以 $\dfrac{\beta}{1+\lambda_2} \geq 0$，所以 $\dfrac{1}{1+\lambda_2}(v_i - \lambda_1) \leq 0$

综上，当 $v_i \geqslant 0$ 时，$w_i = \max(0, \dfrac{1}{1+\lambda_2}(v_i - \lambda_1))$

(2)当 $v_i < 0$ 时：

由于 $w_i v_i \ge 0$，所以 w_i≤0，相对于在 $\underset{w_i}{\operatorname{argmin}}\left(\frac{1}{2}(w_i - v_i)^2 + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2\right)$ 中引入了不等式的约束条件 w_i≤0，引入拉格朗日乘子 β≥0，由 KKT 条件(参考备注)有：

$$\begin{cases} \frac{\partial}{\partial w_i}\left(\frac{1}{2}(w_i - v_i)^2 + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2 + \beta w_i\right) = 0 \\ \beta w_i = 0 \end{cases}$$

所以：$w_i = \frac{1}{1+\lambda_2}(v_i + \lambda_1 - \beta)$，分两种情况：

a.当 w_i<0 时：

由于 βw_i=0，所以 β=0，$w_i = \frac{1}{1+\lambda_2}(v_i + \lambda_1) < 0$，即：$-\frac{1}{1+\lambda_2}(v_i + \lambda_1) > 0$

b.当 w_i=0 时：

由于 β≥0，所以 $\frac{\beta}{1+\lambda_2} \ge 0$，所以 $\frac{1}{1+\lambda_2}(v_i + \lambda_1) \ge 0$，即：$-\frac{1}{1+\lambda_2}(v_i + \lambda_1) \le 0$

综上，当 v_i<0 时，$w_i = \min(0, \frac{1}{1+\lambda_2}(v_i + \lambda_1))$；经化简，当 v_i<0 时，$w_i = -\max(0, -\frac{1}{1+\lambda_2}(v_i + \lambda_1))$。

综合上面两种情况 v_i≥0 和 v_i<0，可以得到 FOBOS 算法在 L1 与 L2 正则化条件下，第 t+1 次第 i 个特征的权重为：

$$\begin{aligned} w_i^{t+1} &= \operatorname{sgn}(v_i)\max(0, \frac{1}{1+\lambda_2}(|v_i| - \lambda_1)) \\ &= \operatorname{sgn}(w_i^t - \eta_i^t g_i^t)\max\{0, \frac{1}{1+\lambda_2}(|w_i^t - \eta_i^t g_i^t| - \lambda_1)\} \end{aligned}$$

通常情况下，λ_2=0，$\lambda_1 = \eta_i^{t+0.5}\lambda$，可以令 $\eta_i^{t+0.5} = f(\frac{1}{\sqrt{t}})$，即一个关于 $\frac{1}{\sqrt{t}}$ 的非增函数。

我们可以发现，在 TG 算法中，令 θ=+∞，α=λ_1，k=1，则第 t+1 次第 i 个特征的权重为：

$$\begin{aligned} w_i^{t+1} &= \begin{cases} \max(0, v_i - \lambda_1) & v_i \ge 0 \\ \max(0, v_i + \lambda_1) & v_i < 0 \end{cases} \\ &= \begin{cases} \max(0, v_i - \lambda_1) & v_i \ge 0 \\ -\max(0, -v_i - \lambda_1) & v_i < 0 \end{cases} \end{aligned}$$

因此在 TG 算法中，$w_i^{t+1} = \operatorname{sgn}(w_i^t - \eta_i^t g_i^t)\max\{0, |w_i^t - \eta_i^t g_i^t| - \lambda_1\}$。所以在 FOBOS 算法中，$\lambda_2$=0 时，与 TG 算法完全一致，故 FOBOS-L1 算法是 TG 算法的一种特例。

3. RDA 算法

除了简单截断法、TG 算法和 FOBOS 算法可以防止过拟合和提高稀疏性，正则对偶平均(RDA, Regularized Dual Averaging)也可以防止过拟合和有效提高特征权重的稀疏性，它是微软 10 年的研究成果，特征的权重向量更新公式为：

$$\boldsymbol{W}^{t+1} = \underset{\boldsymbol{W}}{\operatorname{argmin}}\left\{\frac{1}{t}\sum_{r=1}^{t}\left\langle \boldsymbol{G}^r, \boldsymbol{W}\right\rangle + \Psi(\boldsymbol{W}) + h(\boldsymbol{W})\right\}$$

上式中，$\frac{1}{t}\sum_{r=1}^{t}\left\langle \boldsymbol{G}^r, \boldsymbol{W}\right\rangle$ 表示前面 t 次梯度的平均值与权重 $\boldsymbol{W}$ 的乘积，$\Psi(\boldsymbol{W})$为 L1 正则化与 L2

正则化，$h(W)$ 表示额外正则项，它是一个严格的凸函数（可以是 L2 正则化）。将上式拆分为第 i 个特征维度上，即：

$$w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left\{\bar{g}_i^t w_i + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2 + \frac{1}{2}\gamma w_i^2\right\}$$

其中，$\bar{g}_i^t = \frac{1}{t}\sum_{r=1}^{t} g_i^r$ 。所以对上式右边求导后为 0，即：

$$(\lambda_2+\gamma)w_i + \bar{g}_i^t + \lambda_1\partial|w_i| = 0$$

因此分三种情况 $w_i>0$，$w_i<0$，$w_i=0$ 进行讨论：

(1)当 $w_i>0$ 时：

$$(\lambda_2+\gamma)w_i + \bar{g}_i^t + \lambda_1 = 0$$

所以，$w_i = -\frac{1}{\lambda_2+\gamma}(\bar{g}_i^t + \lambda_1)$，此时 $\bar{g}_i^t < -\lambda_1$；

(2)当 $w_i<0$ 时：

$$(\lambda_2+\gamma)w_i + \bar{g}_i^t - \lambda_1 = 0$$

所以，$w_i = -\frac{1}{\lambda_2+\gamma}(\bar{g}_i^t - \lambda_1)$，此时 $\bar{g}_i^t > \lambda_1$；

(3)当 $w_i=0$ 时，为上面两种情况的反面，即：$|\bar{g}_i^t| \le \lambda_1$；

综合上面的三种，可以得到 RDA 算法在 L1 与 L2 正则化条件下，第 t+1 次第 i 个特征的权重为：

$$w_i^{t+1} = \begin{cases} 0 & |\bar{g}_i^t| \le \lambda_1 \\ -\frac{1}{\lambda_2+\gamma}(\bar{g}_i^t - \operatorname{sgn}(\bar{g}_i^t)\lambda_1) & otherwise \end{cases}$$

这里我们发现，当某个维度上累积梯度平均值的绝对值 $\bar{g}_i^t$ 小于阈值 λ_1 的时候，该维度权重将被置 0，特征权重的稀疏性由此产生

4. FTRL 算法

通过上面，可以知道 FOBOS 算法第 t+1 次第 i 个特征维度权重更新公式为：

$$w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left(\frac{1}{2}(w_i - w_i^t + \eta_i^t g_i^t)^2 + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2\right)$$

等价于(可以适当调整 λ_1 和 λ_2，如上式中 $\lambda_1 = 2\eta_i^t\lambda_1$，$\lambda_2 = 2\eta_i^t\lambda_2$)：

$$w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left(g_i^t w_i + \frac{1}{2\eta_i^t}(w_i - w_i^t)^2 + \frac{1}{2}\eta_i^t(g_i^t)^2 - g_i^t w_i^t + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2\right)$$

由于 $\frac{1}{2}\eta_i^t(g_i^t)^2 - g_i^t w_i^t$ 与 w_i 无关，所以上式等价于：

$$w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left(g_i^t w_i + \frac{1}{2\eta_i^t}(w_i - w_i^t)^2 + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2\right)$$

同时，通过上面可以知道 RDA 算法第 t+1 次第 i 个特征维度权重更新公式为：

$$w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left\{\frac{1}{t}\sum_{r=1}^{t} g_i^r w_i + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2 + \frac{1}{2}\gamma w_i^2\right\}$$

令 $\gamma = \frac{t}{\eta_i^t}$ (可以适当调整 λ_1 和 λ_2，如上式中 $\lambda_1 = t\lambda_1$， $\lambda_2 = t\lambda_2$)，所以上式等价于：

$$w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left\{\sum_{r=1}^{t} g_i^r w_i + \frac{1}{2\eta_i^t}(w_i - 0)^2 + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2\right\}$$

综上，FOBOS 算法和 RDA 算法在第 t+1 次第 i 个特征维度权重更新公式分别为：

$$\begin{cases} w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left(g_i^t w_i + \frac{1}{2\eta_i^t}(w_i - w_i^t)^2 + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2\right) & \text{FOBOS算法} \\ w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left\{\sum_{r=1}^{t} g_i^r w_i + \frac{1}{2\eta_i^t}(w_i - 0)^2 + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2\right\} & \text{RDA算法} \end{cases}$$

可以发现，FOBOS 算法考虑的是当前梯度的影响，RDA 算法则考虑了累积影响；FOBOS 限制 w_i^{t+1} 不能离 w_i^t 太远，而 RDA 算法的 w_i^{t+1} 则不能离 0 太远，因此后者更容易产生稀疏性。

FTRL(Follow the Regularized Leader)算法综合考虑了 FOBOS 算法和 RDA 算法，因此 FTRL 算法第 t+1 次第 i 个特征维度权重更新公式为：

$$w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left\{\sum_{s=1}^{t} g_i^s w_i + \frac{1}{2}\sum_{s=1}^{t}\sigma_i^s(w_i - w_i^s)^2 + \lambda_1|w_i| + \frac{1}{2}\lambda_2 w_i^2\right\}$$

其中， $\frac{1}{\eta_i^t} = \sum_{s=1}^{t}\sigma_i^s$ ， $\sigma_i^t = \frac{1}{\eta_i^t} - \frac{1}{\eta_i^{t-1}}$ ，将上式拆分，等价于：

$$w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left\{\frac{1}{2}\left(\lambda_2 + \sum_{s=1}^{t}\sigma_i^s\right)w_i^2 + \left(\sum_{s=1}^{t} g_i^s - \sum_{s=1}^{t}\sigma_i^s w_i^s\right)w_i + \lambda_1|w_i| + \frac{1}{2}\sum_{s=1}^{t}\sigma_i^s(w_i^s)^2\right\}$$

由于 $\frac{1}{2}\sum_{s=1}^{t}\sigma_i^s(w_i^s)^2$ 相对于 w_i 是常量，所以上式等价于：

$$w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left\{\frac{1}{2}\left(\lambda_2 + \sum_{s=1}^{t}\sigma_i^s\right)w_i^2 + \left(\sum_{s=1}^{t} g_i^s - \sum_{s=1}^{t}\sigma_i^s w_i^s\right)w_i + \lambda_1|w_i|\right\}$$

令 $z_i^t = \sum_{s=1}^{t} g_i^s - \sum_{s=1}^{t}\sigma_i^s w_i^s$ ，所以

$$w_i^{t+1} = \underset{w_i}{\operatorname{argmin}}\left\{\frac{1}{2}\left(\lambda_2 + \sum_{s=1}^{t}\sigma_i^s\right)w_i^2 + z_i^t w_i + \lambda_1|w_i|\right\}$$

所以对上式右边求导后为 0，即：

$$\left(\lambda_2 + \sum_{s=1}^{t}\sigma_i^s\right)w_i + z_i^t + \lambda_1\partial|w_i| = 0$$

因此分三种情况 $w_i>0$，$w_i<0$，$w_i=0$ 进行讨论：

(1)当 $w_i>0$ 时：

$$\left(\lambda_2 + \sum_{s=1}^{t}\sigma_i^s\right)w_i + z_i^t + \lambda_1 = 0$$

所以， $w_i = -\left(\lambda_2 + \sum_{s=1}^{t}\sigma_i^s\right)^{-1}(z_i^t + \lambda_1)$，此时 $z_i^t < -\lambda_1$；

(2)当 $w_i<0$ 时：

$$\left(\lambda_2 + \sum_{s=1}^{t}\sigma_i^s\right)w_i + z_i^t - \lambda_1 = 0$$

所以， $w_i = -\left(\lambda_2 + \sum_{s=1}^{t}\sigma_i^s\right)^{-1}(z_i^t - \lambda_1)$，此时 $z_i^t > \lambda_1$；

(3)当 $w_i=0$ 时，为上面两种情况的反面，即：$|z_i^t| \le \lambda_1$；

综上，第 t+1 次的第 i 个特征权重的更新方程为：

$$w_i^{t+1} = \begin{cases} 0 & |z_i^t| \le \lambda_1 \\ -\left(\lambda_2 + \sum_{s=1}^{t}\sigma_i^s\right)^{-1}\left(z_i^t - \operatorname{sgn}(z_i^t)\lambda_1\right) & otherwise \end{cases}$$

从上面可以看出，当 $\lambda_1=0$，$\lambda_2=0$， $\eta_i^t = \eta$(常数) 时，

$$\begin{aligned} w_i^{t+1} &= -\eta z_i^t = -\eta\left(\sum_{s=1}^{t} g_i^s - \sum_{s=1}^{t}\sigma_i^s w_i^s\right) \\ &= -\eta\left(\sum_{s=1}^{t-1} g_i^s - \sum_{s=1}^{t-1}\sigma_i^s w_i^s + g_i^t - \sigma_i^t w_i^t\right) \\ &= -\eta\left(z_i^{t-1} + g_i^t - \sigma_i^t w_i^t\right) \\ &= -\eta\left(z_i^{t-1} + g_i^t - \left(\frac{1}{\eta} - \frac{1}{\eta}\right)w_i^t\right) \\ &= -\eta z_i^{t-1} - \eta g_i^t \\ &= w_i^t - \eta g_i^t \end{aligned}$$

所以 FTRL 算法就是普通逻辑回归的一种特殊形式。

由于第 t 次第 i 个特征的学习率 η_i^t 为

$$\eta_i^t = \frac{1}{\sqrt{t}}$$

也就是说对于样本的每个特征学习率都是一样的，但是在实际样本中，由于样本每个特征的分布不同，每个特征的学习率也该不同，所以第 t 次第 i 个特征的学习率 η_i^t 为

$$\eta_i^t = \frac{\alpha}{\beta + \sqrt{\sum_{s=1}^{t} g_{s,i}^2}}$$

上式中，α 和 β 都是超参数。

下面来介绍 FTRL 算法的迭代步骤，在介绍迭代以前，首先介绍逻辑回归中样本第 t 次第 i 个特征的梯度 g_i^t 为

$$g_i^t = (h_{\boldsymbol{w}}(\boldsymbol{x}) - y)x_i$$

其中，$\boldsymbol{x}$ 是样本各个维度的特征值，y 是样本标签，x_i 样本第 i 个维度的特征值。同时，

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}}}$$

$$n_i^{t+1} = n_i^t + (g_i^{t+1})^2$$

因此，FTRL 算法的迭代步骤为：

步骤 1：初始化参数，如：w_i^0，α=0.1，β=1，λ_1=0.8，λ_2=0.2，$z_i^0 = 0$，$n_i^0 = 0$

步骤 2：根据第 t 次训练的结果更新 t+1 次第 i 个特征权重的训练结果，即

$$w_i^{t+1} = \begin{cases} 0 & |z_i^t| \le \lambda_1 \\ -(\lambda_2 + \dfrac{\beta + \sqrt{n_i^t}}{\alpha})^{-1}[z_i^t - \operatorname{sgn}(z_i^t)\lambda_1] & \text{其他} \end{cases}$$

步骤 3：对于样本的每一个维度，更新如下参数：

$$g_i^{t+1} = (h_{\boldsymbol{w}}(\boldsymbol{x}) - y)x_i$$

$$n_i^{t+1} = n_i^t + (g_i^{t+1})^2$$

$$z_i^{t+1} = z_i^t + g_i^{t+1} - \frac{1}{\alpha}(\sqrt{n_i^{t+1}} - \sqrt{n_i^t})w_i^{t+1}$$

注：

在含有不等式约束的优化问题中，常用 KKT (Karush-Kuhn-Tucker)条件求解约束优化问题，其中它的前提条件是目标函数是凸优化函数，假设不等式约束优化问题为：

$$\min_x f(x)$$
$$s.t. \quad g(x) \le 0$$

定义拉格朗日函数 $L(x,\beta)$，即：

$$L(x,\beta) = f(x) + \beta g(x)$$

因此，最优解 x 满足 KKT 条件，即：

$$\begin{cases} \dfrac{\partial L(x,\beta)}{\partial x} = 0 \\ \beta g(x) = 0 \\ g(x) \le 0 \\ \beta \ge 0 \end{cases}$$