Adaboost

2. 前向分步算法

2.1 前向分步算法的过程

对于一个加法模型$f(\mathbf{x})=\sum_{m=1}^{M}\beta_m b(\mathbf{x};\gamma_m)$，由M个基函数$b$，通过系数$\beta$线性相加而成，基函数$b$输入$\mathbf{x}$，参数是$\gamma_m$。

对于加法模型$f(\mathbf{x})$而言，可以学习的参数有β_m和γ_m（这两个参数其实是外层和内层的关系）。定义$f(\mathbf{x})$的损失函数为$L(y,f(\mathbf{x}))$，那么学习$f(\mathbf{x})$的参数的过程其实是最小化损失函数

$$\min_{\beta_m,\gamma_m}\sum_{i=1}^{N}L\left(y^{(i)},f(\mathbf{x})\right)=\min_{\beta_m,\gamma_m}\sum_{i=1}^{N}L\left(y^{(i)},\sum_{m=1}^{M}\beta_m b(x^{(i)};\gamma_m)\right)$$

要直接优化这个问题比较复杂。相对的，前向分步算法，通过从前向后，每一步只学习一个基函数及其系数，逐步逼近优化目标函数，从而简化求解复杂度。这样就把内层的$\sum_{m=1}^{M}$去掉，从而变成

$$\min_{\beta,\gamma}\sum_{i=1}^{N}L(y^{(i)},\beta b(x^{(i)};\gamma))$$

通过调整每步的β和γ从而最小化当前步骤的损失函数。这样，前向分步算法将同时求解从$m=1$到M所有的参数β和γ_m的优化问题，变成了逐层求解β和γ的优化问题。

具体来看，第m层的优化问题为

$$\arg\min_{\beta,\gamma}\sum_{i=1}^{N}L(y^{(i)},f_{m-1}(\mathbf{x}^{(i)})+\beta b(\mathbf{x}^{(i)};\gamma))$$

注意到，第m层计算数据点类别的函数，除了包含第m层的基函数以外，还要包括前$m-1$层训练好的加法模型$f_{m-1}(\mathbf{x}^{(i)})$。求解这个$\min$方法得到第$m$层的$(\beta_m^\star,\gamma_m^\star)$，并且更新加法模型

$$f_m(\mathbf{x}^{(i)})=f_{m-1}(\mathbf{x}^{(i)})+\beta_m^\star b(\mathbf{x}^{(i)};\gamma_m^\star)$$

求解$1\cdots M$层的最优参数，然后线性相加，得到最后的最优加法模型$f(\mathbf{x})=f_M(\mathbf{x})=\sum_{m=1}^{M}\beta_m^\star b(x;\gamma_m^\star)$

2.2 前向分步算法与Adaboost算法的关系

当前向分步算法的损失函数是指数函数时，学习的过程等价于Adaboost算法。 假设前向分步算法的基函数用$G(\mathbf{x})$表示，系数用α_m表示。则加法模型可以表示成下面的式子。

$$f(\mathbf{x})=\sum_{m=1}^{M}\alpha_m G_m(\mathbf{x})$$

假设经过$m-1$轮迭代得到了$f_{m-1}(\mathbf{x})$，

$$
\begin{aligned}
f_{m-1}(\mathbf{x}) &= f_{m-2}(\mathbf{x}) + \alpha_{m-1}^{\star} G_{m-1}(\mathbf{x}) \\
&= \alpha_1^{\star} G_1(\mathbf{x}) + \alpha_2^{\star} G_2(\mathbf{x}) + \cdots + \alpha_{m-1}^{\star} G_{m-1}(\mathbf{x})
\end{aligned}
$$

假设在第M轮迭代中，得到了最优的参数$\alpha_M^{\star}$和分类器$G_m(\mathbf{x})$，同时 $f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \alpha_m^{\star} G_m(\mathbf{x})$

为了得到$\alpha_M^{\star}$和分类器$G_m(\mathbf{x})$，必须求解以下目标函数

$$
\begin{aligned}
&\arg\min_{\alpha,G} \sum_{i=1}^{N} \exp\left[-y^{(i)}\left(f_{m-1}(\mathbf{x}^{(i)}) + \alpha G(\mathbf{x}^{(i)})\right)\right] \\
&= \sum_{i=1}^{N} \exp\left[-y^{(i)} f_{m-1}(\mathbf{x}^{(i)}) - y^{(i)} \alpha G(\mathbf{x}^{(i)})\right] \\
&= \sum_{i=1}^{N} \exp\left(-y^{(i)} f_{m-1}(\mathbf{x}^{(i)})\right) \exp\left(-y^{(i)} \alpha G(\mathbf{x}^{(i)})\right) \\
&= \sum_{i=1}^{N} \overline{w}_m^{(i)} \exp[-y^{(i)} \alpha G(\mathbf{x}^{(i)})] \quad \text{定义 } \overline{w}_m^{(i)} = \exp\left(-y^{(i)} f_{m-1}(\mathbf{x}^{(i)})\right)
\end{aligned}
$$

求解这个关于α和$G_m(\mathbf{x})$的最小化问题，可以分两步走，先对$G_m(\mathbf{x})$求最小化，再对α求最小化。

在第一步对$G_m(\mathbf{x})$求最小化，α可以是任意值，但是因为α在$\exp[-y^{(i)} \alpha G(\mathbf{x}^{(i)})]$中，为了不影响幂的符号，这里定义$\alpha > 0$

把所有可以不考虑的项去掉后，目标函数变成

$$
\arg\min_{G} \sum_{i=1}^{N} \overline{w}_m^{(i)} \exp\left(-y^{(i)} G(\mathbf{x}^{(i)})\right)
$$

注意到基本分类器$G_m(\mathbf{x}) : \mathcal{X} \to \{-1, +1\}$，所以当$y^{(i)} = G(\mathbf{x}^{(i)})$时，$-y^{(i)} G(\mathbf{x}^{(i)})) = -1$；当$y^{(i)} \neq G(\mathbf{x}^{(i)})$时，$-y^{(i)} G(\mathbf{x}^{(i)})) = 1$。为了使得这个求和公式得到的结果最小，因为exp是一个单调增函数，所以要尽量使得$y^{(i)} \neq G(\mathbf{x}^{(i)})$的情况越少越好。这也很直观，其实就是使得基本分类器$G(\mathbf{x})$要尽可能不分错。

所以优化以下式子等价于对$G_m(\mathbf{x})$求最小化，从而得到$G_m^{\star}(\mathbf{x})$。

$$
G_m^{\star}(\mathbf{x}) = \arg\min_{G} \sum_{i=1}^{N} \overline{w}_m^{(i)} I(y^{(i)} \neq G(\mathbf{x}^{(i)}))
$$

第二步针对α进行最小化。

$$
\begin{aligned}
&\sum_{i=1}^{N} \overline{w}_m^{(i)} \exp[-y^{(i)} \alpha G(\mathbf{x}^{(i)})] \\
&= \sum_{i;y^{(i)} = G(\mathbf{x}^{(i)})} \overline{w}_m^{(i)} \exp(-\alpha) + \sum_{i;y^{(i)} \neq G(\mathbf{x}^{(i)})} \overline{w}_m^{(i)} \exp(\alpha) \\
&= \exp(-\alpha) \left(\sum_{i=1}^{N} \overline{w}_m^{(i)} - \sum_{i;y^{(i)} \neq G(\mathbf{x}^{(i)})} \overline{w}_m^{(i)} \right) + \exp(\alpha) \sum_{i;y^{(i)} \neq G(\mathbf{x}^{(i)})} \overline{w}_m^{(i)} \\
&= (\exp(\alpha) - \exp(-\alpha)) \sum_{i;y^{(i)} \neq G(\mathbf{x}^{(i)})} \overline{w}_m^{(i)} + \exp(-\alpha) \sum_{i=1}^{N} \overline{w}_m^{(i)}
\end{aligned}
$$

用$G_m^{\star}(\mathbf{x})$代入，对α求偏导数，得到

$$(e^{\alpha}+e^{-\alpha})\sum_{i;y^{(i)}\neq G(\mathbf{x}^{(i)})}\overline{w}_m^{(i)}-e^{-\alpha}\sum_{i=1}^{N}\overline{w}_m^{(i)}$$

令偏导数等于0，整理后得到

$$e^{\alpha}\sum_{i;y^{(i)}\neq G(\mathbf{x}^{(i)})}\overline{w}_m^{(i)}=e^{-\alpha}\left(\sum_{i=1}^{N}\overline{w}_m^{(i)}-\sum_{i;y^{(i)}\neq G(\mathbf{x}^{(i)})}\overline{w}_m^{(i)}\right)=e^{-\alpha}\sum_{i;y^{(i)}=G(\mathbf{x}^{(i)})}\overline{w}_m^{(i)}$$

把α整理到一边可得

$$e^{2\alpha}=\frac{\sum_{i;y^{(i)}=G(\mathbf{x}^{(i)})}\overline{w}_m^{(i)}}{\sum_{i;y^{(i)}\neq G(\mathbf{x}^{(i)})}\overline{w}_m^{(i)}}$$

得到

$$\alpha=\frac{1}{2}\log\frac{\sum_{i;y^{(i)}=G(\mathbf{x}^{(i)})}\overline{w}_m^{(i)}}{\sum_{i;y^{(i)}\neq G(\mathbf{x}^{(i)})}\overline{w}_m^{(i)}}$$

定义误差率

$$e_m=\frac{\sum_{i;y^{(i)}\neq G(\mathbf{x}^{(i)})}\overline{w}_m^{(i)}}{\sum_{i=1}^{N}\overline{w}_m^{(i)}}$$

代入α中可得

$$\alpha=\frac{1}{2}\log\frac{1-e_m}{e_m}$$

这里的α就是求得的最优$\alpha_m^{\star}$

根据加法模型的定义，$f_m(\mathbf{x})=f_{m-1}(\mathbf{x})+\alpha_m G_m(\mathbf{x})$，所以$f_{m-1}(\mathbf{x})=f_m(\mathbf{x})-\alpha_m G_m(\mathbf{x})$，将其带入$\overline{w}_m$的表达式有

$$\begin{aligned}\overline{w}_m&=\exp\left(-yf_{m-1}(\mathbf{x})\right)\\&=e^{-y(f_m(\mathbf{x})-\alpha_m G_m(\mathbf{x}))}\\&=e^{-yf_m(\mathbf{x})+y\alpha_m G_m(\mathbf{x})}\\&=\overline{w}_{m+1}e^{y\alpha_m G_m(\mathbf{x})}\end{aligned}$$

整理得到

$$\overline{w}_{m+1}=\overline{w}_m e^{-y\alpha_m G_m(\mathbf{x})}$$

可以看到，误差率e_m，基函数$G_m\mathbf{x}$的系数α_m，还有权值更新过程$\overline{w}_{m+1}$都和1.1节中Adaboost算法介绍的一致，因此对于前向分步算法而言，当损失函数为指数函数时，等价于Adaboost算法。