AdaBoost

本文介绍Boosting算法及一些代表算法。准确来说，Boosting算法更多的是一种思想。例如一个分类任务，如果训练一个分类器可以做到60%的正确率。那么同时训练多个分类器，利用投票的方法来对数据集进行分类，经验上可以获得更高的正确率。这就是Boosting的思想。

1. AdaBoost算法

给定一组训练数据集

$$T = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \cdots, (\mathbf{x}^{(N)}, y^{(N)})\}$$

该数据集包含N个样本点。每个样本点的$\mathbf{x}$加粗显示，表示这是一个向量，$\mathbf{x} \in \mathbb{R}^n$，当然如果n=1，则$\mathbf{x}$是一个标量。

在$\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)})$中的每一个维度，表示该样本点的一个特征，样本集中的每个样本点有n个维度或特征。

$y^{(i)}$表示第i个样本点的类别，$y \in \{+1, -1\}$，当$y^{(i)} = 1$，则表示$\mathbf{x}^{(i)}$是正例。当$y^{(i)} = -1$，则表示$\mathbf{x}^{(i)}$是负例。

Adaboost算法从数据集中学习一系列弱分类器或基本分类器，并且线性组成一个强分类器。

1.1 Adaboost算法的过程

1. 初始化训练数据的权值分布为等权

$$D_1 = (w_1^{(1)}, w_1^{(2)}, \cdots, w_1^{(N)}), \quad w_1^{(i)} = \frac{1}{N}, i = 1, 2, \cdots N$$

其中D的下标表示第几次的权值分布。可以看到，第一次每个数据的权值初始化为等权。

2. 对$m = 1, 2, \ldots, M$进行循环
   - 使用权值分布为D_m的训练数据训练一个基本分类器$G_m(\mathbf{x}) : \mathcal{X} \rightarrow \{-1, +1\}$
   - 计算$G_m(\mathbf{x})$在训练数据上的分类误差率

$$e_m = P(G_m(\mathbf{x}^{(i)}) \neq y^{(i)}) = \sum_{i=1}^{N} w_m^{(i)} I(G_m(\mathbf{x}^{(i)}) \neq y^{(i)})$$

   - 计算$G_m(\mathbf{x})$的系数

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

   - 更新训练数据集的权值分布$D_{m+1} = (w_{m+1}^{(1)}, w_{m+1}^{(2)}, \cdots, w_{m+1}^{(N)})$

$$w_{m+1}^{(i)} = \frac{w_m^{(i)}}{Z_m} \exp(-\alpha_m y^{(i)} G_m(\mathbf{x}^{(i)})), \quad i = 1, 2, \cdots, N$$

   其中Z_m是规范化因子，为了使得D_{m+1}也是一个概率分布。

$$Z_m = \sum_{i=1}^{N} w_m^{(i)} \exp(-\alpha_m y^{(i)} G_m(x^{(i)}))$$

3. 构建基本分类器的线性组合

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

最终的分类器为

$$G(\mathbf{x}) = \mathrm{sign}(f(\mathbf{x})) = \mathrm{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(\mathbf{x})\right)$$

1.2 Adaboost算法的说明

1. 假设训练数据集具有均匀的权值分布，即在训练$G_1(\mathbf{x})$的时候，每个数据都一样重要。
2. 在之后的训练中，权值D_m要根据第$m-1$次训练得到的分类器$G_{m-1}(\mathbf{x})$在数据集上的误差e_{m-1}作出调整。利用D_m训练$G_m(\mathbf{x})$。
3. 根据误差率的计算公式

$$\begin{aligned} e_m &= P(G_m(\mathbf{x}^{(i)}) \neq y^{(i)}) = \sum_{i=1}^{N} w_m^{(i)} I(G_m(\mathbf{x}^{(i)}) \neq y^{(i)}) \\ &= \sum_{G_m(\mathbf{x}^{(i)}) \neq y^{(i)}}^{N} w_m^{(i)} \end{aligned}$$

注意到每轮在计算权值$w_m^{(i)}$的时候，都要进行规范化操作，所以每轮的权值都满足$\sum_{i=1}^{N} w_m^{(i)} = 1$，误差率是把其中$G_m(\mathbf{x})$分错的样本点的权值求和，其值要小于等于1。

4. 根据基本分类器的系数α_m的计算公式可以看出，当$G_m(\mathbf{x})$的误差率$e_m \leq 0.5$，分子大于分母，分式大于1，则$\alpha_m \geq 0$。当e_m越小，α_m越大，表明第m轮的基本分类器$G_m(\mathbf{x})$越重要。
5. 更新权值过程中，要看分类器是否正确分类样本点。

$$w_{m+1}^{(i)} = \begin{cases} \frac{w_m^{(i)}}{Z} e^{(-\alpha_m)}, & G_m(\mathbf{x}^{(i)}) = y^{(i)} \\ \frac{w_m^{(i)}}{Z} e^{(\alpha_m)}, & G_m(\mathbf{x}^{(i)}) \neq y^{(i)} \end{cases}$$

当$\alpha_m > 0$，即$G_m(\mathbf{x})$表现不错的轮次，当$G_m(\mathbf{x}^{(i)}) = y^{(i)}$时，$e^{(-\alpha_m)} < 1$，$w_{m+1}^{(i)}$会被调低。而当$G_m(\mathbf{x}^{(i)}) \neq y^{(i)}$时，$e^{(\alpha_m)} > 1$，$w_{m+1}^{(i)}$会被调高。说明会重视分类错误的点，轻视分类正确的点。

当$\alpha_m < 0$，即$G_m(\mathbf{x})$表现不好的轮次，当$G_m(\mathbf{x}^{(i)}) = y^{(i)}$时，$e^{(-\alpha_m)} > 1$，$w_{m+1}^{(i)}$会被调低。而当$G_m(\mathbf{x}^{(i)}) \neq y^{(i)}$时，$e^{(\alpha_m)} < 1$，$w_{m+1}^{(i)}$会被调低。说明会重视分类正确的点，轻视分类错误的点。

6. 在对多轮的基本分类器利用α_m进行加权得到$f(\mathbf{x})$的过程中，加权系数α_m并没有要求合等于1。同时$f(\mathbf{x})$的符号表示类别，绝对值表示确信度。

1.3 Adaboost训练误差上界

$$\frac{1}{N}\sum_{i=1}^{N} I(G(\mathbf{x}^{(i)}) \neq y^{(i)}) \leq \frac{1}{N}\sum_{i=1}^{N} \exp(-y^{(i)} f(\mathbf{x}^{(i)})) = \prod_{m=1}^{M} Z_m$$

先证明左边的不等式。

当$G(\mathbf{x}^{(i)}) \neq y^{(i)}$时，$I(G(\mathbf{x}^{(i)}) \neq y^{(i)}) = 1$，而$-y^{(i)} f(\mathbf{x}^{(i)}) \geq 0$，所以$\exp(-y^{(i)} f(\mathbf{x}^{(i)})) \geq 1$。

当$G(\mathbf{x}^{(i)}) = y^{(i)}$，$I(G(\mathbf{x}^{(i)}) \neq y^{(i)}) = 0$，而$-y^{(i)} f(\mathbf{x}^{(i)}) \leq 0$，所以$0 \leq \exp(-y^{(i)} f(\mathbf{x}^{(i)})) \leq 1$。所以无论哪种情况$\exp(-y^{(i)} f(\mathbf{x}^{(i)})) \geq I(G(\mathbf{x}^{(i)}) \neq y^{(i)})$，所以左边的不等式得证。

接下来证明右边的等式。

首先注意到第一轮的权值

$$w_1^{(i)} = \frac{1}{N}$$

所以可以把权值$w_1^{(i)}$放入$\sum_{i=1}^{N}$中得到

$$\sum_{i=1}^{N} w_1^{(i)} \exp(-y^{(i)} f(\mathbf{x}^{(i)}))$$

同时将$f(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m G_m(\mathbf{x})$代入得到

$$\begin{aligned} \sum_{i=1}^{N} w_1^{(i)} \exp(-y^{(i)} \sum_{m=1}^{M} \alpha_m G_m(\mathbf{x}^{(i)})) &= \sum_{i=1}^{N} w_1^{(i)} \exp(-\sum_{m=1}^{M} y^{(i)} \alpha_m G_m(\mathbf{x}^{(i)})) \\ &= \sum_{i=1}^{N} w_1^{(i)} \prod_{m=1}^{M} \exp(-y^{(i)} \alpha_m G_m(\mathbf{x}^{(i)})) \end{aligned} \tag{1}$$

根据

$$w_{m+1}^{(i)} = \frac{w_m^{(i)}}{Z_m} \exp(-\alpha_m y^{(i)} G_m(\mathbf{x}^{(i)})), \quad i = 1, 2, \cdots, N$$

可知

$$w_{m+1}^{(i)} Z_m = w_m^{(i)} \exp(-\alpha_m y^{(i)} G_m(\mathbf{x}^{(i)})), \quad i = 1, 2, \cdots, N$$

所以

$$w_2^{(i)} Z_1 = w_1^{(i)} \exp(-\alpha_1 y^{(i)} G_1(\mathbf{x}^{(i)})), \quad i = 1, 2, \cdots, N$$

代入(1)式中，可得

$$
\begin{aligned}
&\sum_{i=1}^{N} w_1^{(i)} \prod_{m=1}^{M} \exp(-y^{(i)} \alpha_m G_m(\mathbf{x}^{(i)})) \\
&= \sum_{i=1}^{N} w_1^{(i)} \exp(-y^{(i)} \alpha_1 G_1(\mathbf{x}^{(i)})) \prod_{m=2}^{M} \exp(-y^{(i)} \alpha_m G_m(\mathbf{x}^{(i)})) \\
&= \sum_{i=1}^{N} w_2^{(i)} Z_1 \prod_{m=2}^{M} \exp(-y^{(i)} \alpha_m G_m(\mathbf{x}^{(i)})) \\
&= \sum_{i=1}^{N} Z_1 w_3^{(i)} Z_2 \prod_{m=3}^{M} \exp(-y^{(i)} \alpha_m G_m(\mathbf{x}^{(i)})) \\
&= \cdots \\
&= \sum_{i=1}^{N} Z_1 Z_2 \cdots Z_{M-1} w_M^{(i)} \exp(-y^{(i)} \alpha_M G_M(\mathbf{x}^{(i)})) \\
&= \sum_{i=1}^{N} Z_1 Z_2 \cdots Z_{M-1} w_{M+1}^{(i)} Z_M \\
&= \sum_{i=1}^{N} w_{M+1}^{(i)} \prod_{m=1}^{M} Z_m \quad (\text{w求和等于} 1) \\
&= \prod_{m=1}^{M} Z_m
\end{aligned}
$$

右边的等式得证。

可以看到，在每一个轮次中，通过选择适当的$G_m(\mathbf{w})$使得Z_m最小，可以使训练误差下降最快。

1.4 二分类问题Adaboost训练误差上界

在1.3节中的推导可知，Adaboost算法的训练误差上界是$\prod_{m=1}^{M} Z_m$。如果$y^{(i)} \in \{-1, +1\}$，则考虑一轮的Z_m有以下推导

$$
\begin{aligned}
Z_m &= \sum_{i=1}^{N} w_m^{(i)} \exp(-\alpha_m y^{(i)} G_m(x^{(i)})) \\
&= \sum_{i, y^{(i)} \neq G_m(x^{(i)})}^{N} w_m^{(i)} \exp(\alpha_m) + \sum_{i, y^{(i)} = G_m(x^{(i)})}^{N} w_m^{(i)} \exp(-\alpha_m) \\
&= \sum_{i, y^{(i)} \neq G_m(x^{(i)})}^{N} w_m^{(i)} \exp(\frac{1}{2} \log \frac{1-e_m}{e_m}) + \sum_{i, y^{(i)} = G_m(x^{(i)})}^{N} w_m^{(i)} \exp(-\frac{1}{2} \log \frac{1-e_m}{e_m}) \\
&= \sum_{i, y^{(i)} \neq G_m(x^{(i)})}^{N} w_m^{(i)} \sqrt{\frac{1-e_m}{e_m}} + \sum_{i, y^{(i)} = G_m(x^{(i)})}^{N} w_m^{(i)} \sqrt{\frac{e_m}{1-e_m}} \\
&= e_m \sqrt{\frac{1-e_m}{e_m}} + (1-e_m) \sqrt{\frac{1-e_m}{e_m}} \\
&= 2\sqrt{e_m(1-e_m)} = \sqrt{1-4\gamma_m^2}
\end{aligned}
$$

其中 $\gamma_m = \frac{1}{2} - e_m$

根据泰勒式展开

$$
\begin{aligned}
&(1+x)^a = 1 + \frac{a}{1!}x + \frac{a(a-1)}{2!}x^2 + \frac{a(a-1)(a-2)}{3!}x^3 + o(x^3) \\
&e^x = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + o(x^3)
\end{aligned}
$$

则$\sqrt{1-4\gamma_m^2}$在$x=0$处的泰勒展开式为

$$
\begin{aligned}
\sqrt{1-4\gamma_m^2} &= 1 + \frac{\frac{1}{2}}{1}(-4\gamma_m^2) + \frac{\frac{1}{2}(\frac{1}{2}-1)}{2!}(-4\gamma_m^2)^2 + \frac{\frac{1}{2}(\frac{1}{2}-1)(\frac{1}{2}-2)}{3!}(-4\gamma_m^2)^3 + o(x^3) \\
&= 1 - 2\gamma_m^2 - 2\gamma_m^4 - 4\gamma_m^6 + o(x^3)
\end{aligned}
$$

而

$$
\begin{aligned}
\exp(-2\gamma_m^2) &= 1 + \frac{1}{1}(-2\gamma_m^2) + \frac{1}{2!}(-2\gamma_m^2)^2 + \frac{1}{3!}(-2\gamma_m^2)^3 + o(x^3) \\
&= 1 - 2\gamma_m^2 + 2\gamma_m^4 - \frac{8}{6}\gamma_m^6 + o(x^3)
\end{aligned}
$$

可以看到$\exp(-2\gamma_m^2) \geq \sqrt{1-4\gamma_m^2}$，所以

$$
\prod_{m=1}^{M} Z_m = \prod_{m=1}^{M} \sqrt{1-4\gamma_m^2} \leq \prod_{m=1}^{M} \exp(-2\gamma_m^2) = \exp(-2\sum_{m=1}^{M}\gamma_m^2)
$$

如果存在γ，使得对于每一轮的$\gamma_m \geq \gamma$，有

$$
\exp(-2\sum_{m=1}^{M}\gamma_m^2) \geq \exp(-2\sum_{m=1}^{M}\gamma^2) = \exp(-2M\gamma^2)
$$

结合1.3节和这里的推导可知，错误率的上界是

$$
\frac{1}{N}\sum_{i=1}^{N} I(G(\mathbf{x}^{(i)}) \neq y^{(i)}) \leq \exp(-2M\gamma^2)
$$

表明Adaboost的训练误差是以指数速率下降的。