模型组合方法——Boosting

陈靖

2016-08-11

滴滴

滴滴一下 美好出行

常用的模型组合方法

滴滴
滴滴一下 美好出行

- Averaging
- Voting
- Bagging（RandomForest）
- Boosting（GBDT）
- Stacking
- Blending

为什么要组合模型？
组合模型一定能提升效果吗？

泛化误差=偏差+方差+噪声

完全独立不相关 $Var\left(\frac{\sum X_i}{n}\right)=\frac{Var(X_i)}{n}$

完全相关 $Var\left(\frac{\sum X_i}{n}\right)=Var(X_i)$

目录

- Adaptive Boosting（AdaBoost）
- Gradient Boosting
- 扩展内容xgboost

Error Function

$$E = \frac{1}{N}\sum_{n=1}^{N} err(y_n, h(x_n))$$

$$err = \begin{cases} 0\,, y_n = sign(h(x_n)) \\ 1, y_n \neq sign(h(x_n)) \end{cases}$$

Weighted Error Function

$$E = \frac{1}{N}\sum_{n=1}^{N} u_n \cdot err(y_n, h(x_n))$$

AdaBoost

滴滴
滴滴一下 美好出行

将一系列弱分类器组合为强分类器

$$u^{(1)} = [\frac{1}{N}, \frac{1}{N}, \dots, \frac{1}{N}]$$

for t = 1,2,…,T

obtain h_t to minimize weighted error function

update $u^{(t)}$ to $u^{(t+1)}$: $u_n^{(t+1)} = u_n^{(t)} e^{-y_n \alpha_t h_t(x_n)}$

where $\alpha_t = \ln\left(\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\right), \epsilon_t = \frac{\sum_{n=1}^{N} u_n^{(t)} [\![y_n \neq h_t(x_n)]\!]}{\sum_{n=1}^{N} u_n^{(t)}}$

为什么是这个权重?

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

AdaBoost

滴滴
滴滴一下 美好出行

$$u_n^{(T+1)} = u_n^{(T)} e^{-y_n \alpha_T h_T(x)} = u_n^{(1)} \prod_{t=1}^{T} e^{-y_n \alpha_t h_t(x_n)}$$

$$= \frac{1}{N} e^{\left(-y_n \sum_{t=1}^{T} \alpha_t h_t(x_n)\right)}$$

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

如果$u_n^{(T+1)}$越小，说明$y_n \sum_{t=1}^{T} \alpha_t h_t(x)$越大，说明$\sum_{t=1}^{T} \alpha_t h_t(x)$越趋向于与$y_n$同号

AdaBoost

AdaBoost Error Function

minimizes $$\sum_{n=1}^{N} u_n^{(T+1)} = \frac{1}{N}\sum_{n=1}^{N} e^{\left(-y_n \sum_{t=1}^{T} \alpha_t h_t(x_n)\right)}$$

AdaBoost

滴滴
滴滴一下 美好出行

假设找到一个函数$h(x_n)$，在这个函数上走η的长度

$$\hat{E} = \frac{1}{N}\sum_{n=1}^{N} e^{-y_n\left(\sum_{l=1}^{t-1}\alpha_l h_l(x_n)+\eta h(x_n)\right)}$$

$$= \sum_{n=1}^{N} u_n^{(t)} e^{(-y_n \eta h(x_n))}$$

$$\approx \sum_{n=1}^{N} u_n^{(t)}(1 - y_n \eta h(x_n)) = \sum_{n=1}^{N} u_n^{(t)} + \eta \sum_{n=1}^{N} u_n^{(t)}(-y_n h(x_n))$$

AdaBoost

$$\sum_{n=1}^{N} u_n^{(t)}\left(-y_n h(x_n)\right) = \sum_{n=1}^{N} u_n^{(t)} \begin{cases} -1\,, y_n = h(x_n) \\ 1, y_n \neq h(x_n) \end{cases}$$

$$= -\sum_{n=1}^{N} u_n^{(t)} + \sum_{n=1}^{N} u_n^{(t)} \begin{cases} 0\,, y_n = h(x_n) \\ 2, y_n \neq h(x_n) \end{cases}$$

$$= -\sum_{n=1}^{N} u_n^{(t)} + 2\mathrm{E}(\mathrm{h}) \cdot N$$

$$E = \frac{1}{N}\sum_{n=1}^{N} u_n \cdot err(y_n, h(x_n))$$

AdaBoost

滴滴
滴滴一下 美好出行

寻找最优步长η

$$\hat{E} = \sum_{n=1}^{N} u_n^{(t)} e^{(-y_n \eta h_t(x_n))}$$

$$= \sum_{y_n = h_t(x_n)} u_n^{(t)} e^{-\eta} + \sum_{y_n \neq h_t(x_n)} u_n^{(t)} e^{\eta}$$

$$= \sum_{y_n = h_t(x_n)} u_n^{(t)} e^{-\eta} + \sum_{y_n \neq h_t(x_n)} u_n^{(t)} e^{\eta}$$

似曾相识？

$$\frac{\partial \hat{E}}{\partial \eta} = \eta \left(\sum_{y_n \neq h_t(x_n)} u_n^{(t)} e^{\eta} - \sum_{y_n = h_t(x_n)} u_n^{(t)} e^{-\eta} \right) = 0 \rightarrow \eta = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t} = \ln \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}, \epsilon_t = \frac{\sum_{y_n \neq h_t(x_n)} u_n^{(t)}}{\sum u_n^{(t)}}$$

Gradient Boosting

滴滴
滴滴一下 美好出行

AdaBoost

$$\frac{1}{N}\sum_{n=1}^{N}\exp\left(-y_n\left(\sum_{l=1}^{t-1}\alpha_l h_l(x_n)+\eta h(x_n)\right)\right)$$

GradientBoost

$$\frac{1}{N}\sum_{n=1}^{N} err\left(\sum_{l=1}^{t-1}\alpha_l h_l(x_n)+\eta h(x_n), y_n\right)$$

Gradient Boosting

GradientBoost

$$\frac{1}{N}\sum_{n=1}^{N} err\left(\sum_{l=1}^{t-1}\alpha_l h_l(x_n) + \eta h(x_n), y_n\right)$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} err(s_n, y_n) + \frac{1}{N}\sum_{n=1}^{N} \eta h(x_n) \frac{\partial err(s, y_n)}{s}|_{s=s_n}$$

Gradient Boosting

滴滴 滴滴一下 美好出行

$err = (s - y)^2$

$$= constant + \frac{\eta}{N}\sum_{n=1}^{N} h(x_n) \cdot 2(s_n - y_n)$$

$\Rightarrow$ $h(x_n)$的方向为 $-(s_n - y_n) = (y_n - s_n)$

残差

Gradient Boosting

最优步长η

$$\frac{1}{N}\sum_{n=1}^{N} err(s_n + \eta h_t(x_n), y_n)$$

$$= \frac{1}{N}\sum_{n=1}^{N} (s_n + \eta h_t(x_n) - y_n)^2$$

$$= \frac{1}{N}\sum_{n=1}^{N} \left((y_n - (s_n + \eta h_t(x_n))\right)^2$$

应用

Gradient Boosting+Decision Tree：GBDT

$h_t(x)$

$sign(y-s)$

回归问题：$err=(s-y)^2$或者$|s-y|$

分类问题：$err=\ln\big(1+exp(-ywx)\big)$

⋮

扩展

滴滴
滴滴一下 美好出行

Xgboost

$$Obj^{(t)} = \sum_{i=1}^{N} err\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + constant$$

二阶泰勒展开

$$Obj^{(t)} \approx \sum_{i=1}^{N} \left[err\left(y_i, \hat{y}_i^{(t-1)}\right) + +g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) + constant$$

$$g_i = \frac{\partial err\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}} \qquad h_i = \frac{\partial err\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial^2 \hat{y}_i^{(t-1)}}$$

扩展

Xgboost

$$f_t(x) = w_{q(x)} \qquad I_j = \{i|q(x_i) = j\}$$

$$\begin{aligned} Obj^{(t)} &\simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \tfrac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \\ &= \sum_{i=1}^{n} \left[g_i w_{q(x_i)} + \tfrac{1}{2} h_i w_{q(x_i)}^2\right] + \gamma T + \lambda \tfrac{1}{2} \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \tfrac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2\right] + \gamma T \end{aligned}$$

$$w_j^* = -\frac{G_j}{H_j+\lambda} \qquad Obj = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j+\lambda} + \gamma T$$

$$Gain = \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda} - \gamma$$

扩展

Xgboost特点

- 速度快
- 精度高
- 可定制损失函数

Reference

- Friedman J H. Greedy function approximation: a gradient boosting
- Chen T, Guestrin C. Xgboost: A scalable tree boosting system

滴滴
滴滴一下 美好出行

Q&A，Thanks