深度之眼
deepshare.net

法律声明

课程详情请咨询

- 微信公众号：深度之眼
- 客服微信号：deepshare0920

公众号

微 信

deepshare.net
深度之眼

西瓜书公式推导

导师：Sm1les（Datawhale南瓜书项目负责人）

deepshare.net
深度之眼

决策树公式推导

Derivation of Decision Tree

本节大纲

Outline

先修内容：西瓜书4.1、4.2

1.ID3决策树

2.C4.5决策树

3.CART决策树

ID3决策树

ID3 algorithm

信息熵——度量样本集合纯度最常用的一种指标，其定义如下

$$\operatorname{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

其中，$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$ 表示样本集合，$|\mathcal{Y}|$ 表示样本类别总数，p_k 表示第k类

样本所占的比例，且 $0 \leq p_k \leq 1, \sum_{k=1}^{|\mathcal{Y}|} p_k = 1$。 $\operatorname{Ent}(D)$ 值越小，纯度越高。

ID3决策树

ID3 algorithm

证明：$0 \leq \text{Ent}(D) \leq \log_2 |\mathcal{Y}|$

求 $\text{Ent}(D)$ 最大值：

若令 $|\mathcal{Y}| = n, p_k = x_k$，那么信息熵 $\text{Ent}(D)$ 就可以看作一个n元实值函数，也即

$$\text{Ent}(D) = f(x_1, ..., x_n) = -\sum_{k=1}^{n} x_k \log_2 x_k$$

其中，$0 \leq x_k \leq 1, \sum_{k=1}^{n} x_k = 1$，下面考虑求该多元函数的最值。

ID3决策树

ID3 algorithm

如果不考虑约束 $0 \leq x_k \leq 1$ 仅考虑 $\sum_{k=1}^{n} x_k = 1$ 的话，对 $f(x_1, ..., x_n)$ 求最大值等价于如下最小化问题

$$\begin{aligned} \min \ & \sum_{k=1}^{n} x_k \log_2 x_k \\ \text{s.t.} \ & \sum_{k=1}^{n} x_k = 1 \end{aligned}$$

显然，在 $0 \leq x_k \leq 1$ 时，此问题为凸优化问题，而对于凸优化问题来说，满足KKT条件的点即为最优解。由于此最小化问题仅含等式约束，那么能令其拉格朗日函数的一阶偏导数等于0的点即为满足KKT条件的点。

参考文献:王燕军，梁治安. 最优化基础理论与方法[M]. 复旦大学出版社，2011.

ID3决策树

ID3 algorithm

根据拉格朗日乘子法可知，该优化问题的拉格朗日函数为

$$L(x_1, ..., x_n, \lambda) = \sum_{k=1}^{n} x_k \log_2 x_k + \lambda(\sum_{k=1}^{n} x_k - 1)$$

对拉格朗日函数分别关于 $x_1, ..., x_n, \lambda$ 求一阶偏导数，并令偏导数等于0可得

$$\frac{\partial L(x_1, ..., x_n, \lambda)}{\partial x_1} = \frac{\partial}{\partial x_1}\left[\sum_{k=1}^{n} x_k \log_2 x_k + \lambda(\sum_{k=1}^{n} x_k - 1)\right] = 0$$

$$= \log_2 x_1 + x_1 \cdot \frac{1}{x_1 \ln 2} + \lambda = 0$$

$$= \log_2 x_1 + \frac{1}{\ln 2} + \lambda = 0$$

$$\Rightarrow \lambda = -\log_2 x_1 - \frac{1}{\ln 2}$$

ID3决策树

ID3 algorithm

同理可得

$$\lambda = -\log_2 x_1 - \frac{1}{\ln 2} = -\log_2 x_2 - \frac{1}{\ln 2} = ... = -\log_2 x_n - \frac{1}{\ln 2}$$

又因为

$$\frac{\partial L(x_1, ..., x_n, \lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda}\left[\sum_{k=1}^{n} x_k \log_2 x_k + \lambda(\sum_{k=1}^{n} x_k - 1)\right] = 0$$

$$\Rightarrow \sum_{k=1}^{n} x_k = 1$$

所以可以解得

$$x_1 = x_2 = ... = x_n = \frac{1}{n}$$

ID3决策树

ID3 algorithm

又因为 x_k 还需满足约束 $0 \le x_k \le 1$ 显然 $0 \le \frac{1}{n} \le 1$ 所以 $x_1 = x_2 = ... = x_n = \frac{1}{n}$ 是满足所有约束的最优解，也即为当前最小化问题的最小值点，同时也是 $f(x_1, ..., x_n)$ 的最大值点。将 $x_1 = x_2 = ... = x_n = \frac{1}{n}$ 代入 $f(x_1, ..., x_n)$ 中可得

$$f(\frac{1}{n}, ..., \frac{1}{n}) = -\sum_{k=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = -n \cdot \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n$$

所以 $f(x_1, ..., x_n)$ 在满足约束 $0 \le x_k \le 1, \sum_{k=1}^{n} x_k = 1$ 时的最大值为 $\log_2 n$

ID3决策树

ID3 algorithm

求 $\mathrm{Ent}(D)$ 最小值：

如果不考虑约束 $\sum_{k=1}^{n} x_k = 1$ ，仅考虑$0 \leq x_k \leq 1$ 的话， $f(x_1,...,x_n)$ 可以看做是n个互不相关的一元函数的加和，也即

$$f(x_1,...,x_n) = \sum_{k=1}^{n} g(x_k)$$

其中，$g(x_k) = -x_k \log_2 x_k$，$0 \leq x_k \leq 1$ 那么当 $g(x_1), g(x_2), ..., g(x_n)$ 分别取到其最小值时，$f(x_1,...,x_n)$ 也就取到了最小值。由于$g(x_1), g(x_2), ..., g(x_n)$ 的定义域和函数表达式均相同，所以只需求出 $g(x_1)$ 的最小值也就求出了$g(x_2), ..., g(x_n)$ 的最小值。下面考虑求 $g(x_1)$ 的最小值。首先对$g(x_1)$ 关于x_1求一阶和二阶导数

ID3决策树

ID3 algorithm

$$g'(x_1)=\frac{d(-x_1\log_2 x_1)}{dx_1}=-\log_2 x_1-x_1\cdot\frac{1}{x_1\ln 2}=-\log_2 x_1-\frac{1}{\ln 2}$$

$$g''(x_1)=\frac{d\left(g'(x_1)\right)}{dx_1}=\frac{d\left(-\log_2 x_1-\dfrac{1}{\ln 2}\right)}{dx_1}=-\frac{1}{x_1\ln 2}$$

显然，当 $0\le x_k\le 1$ 时 $g''(x_1)=-\frac{1}{x_1\ln 2}$ 恒小于0，所以 $g(x_1)$ 是一个在其定义域范围内开口向下的凹函数，那么其最小值必然在边界取，于是分别取 $x_1=0$ 和 $x_1=1$，代入 $g(x_1)$ 可得

$$g(0)=-0\log_2 0=0$$

$$g(1)=-1\log_2 1=0$$

ID3决策树

ID3 algorithm

所以，$g(x_1)$ 的最小值为0，同理可得 $g(x_2),...,g(x_n)$的最小值也为0，那么 $f(x_1,...,x_n)$ 的最小值此时也为0。但是，此时是仅考虑 $0 \le x_k \le 1$ 时取到的最小值，若考虑约束 $\sum\limits_{k=1}^{n} x_k = 1$ 的话，那么 $f(x_1,...,x_n)$ 的最小值一定大于等于0。如果令某个 $x_k = 1$，那么根据约束 $\sum\limits_{k=1}^{n} x_k = 1$ 可知 $x_1 = x_2 = ... = x_{k-1} = x_{k+1} = ... = x_n = 0$，将其代入 $f(x_1,...,x_n)$ 可得

$$f(0,0,...,0,1,0,...,0) = -0\log_2 0 - 0\log_2 0... - 0\log_2 0 - 1\log_2 1 - 0\log_2 0... - 0\log_2 0 = 0$$

所以 $x_k = 1, x_1 = x_2 = ... = x_{k-1} = x_{k+1} = ... = x_n = 0$ 一定是 $f(x_1,...,x_n)$ 在满足约束 $0 \le x_k \le 1, \sum\limits_{k=1}^{n} x_k = 1$ 的条件下的最小值点，其最小值为0。

ID3决策树

ID3 algorithm

条件熵——在已知样本属性a的取值情况下，度量样本集合纯度的一种指标

$$H(D|a)=\sum_{v=1}^{V}\frac{|D^v|}{|D|}\operatorname{Ent}\left(D^v\right)$$

其中，a表示样本的某个属性，假定属性a有V个可能的取值$\{a^1,a^2,...,,a^V\}$样本集合D中在属性a上取值为 a^v 的样本记为 D^v，$\operatorname{Ent}(D^v)$ 表示样本集合 D^v 的信息熵。$H(D|a)$ 值越小，纯度越高。

ID3决策树

ID3 algorithm

ID3决策树——以信息增益为准则来选择划分属性的决策树

信息增益:

$$\operatorname{Gain}(D, a) = \operatorname{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Ent}\left(D^v\right)$$

$$= \operatorname{Ent}(D) - H(D|a)$$

选择信息增益值最大的属性作为划分属性，因为信息增益越大，则意味着使用该属性来进行划分所获得的“纯度提升”越大

ID3决策树

ID3 algorithm

以信息增益为划分准则的ID3决策树对可取值数目较多的属性有所偏好

$$
\begin{aligned}
\operatorname{Gain}(D, a) &= \operatorname{Ent}(D)-\sum_{v=1}^{V} \frac{\left|D^{v}\right|}{|D|} \operatorname{Ent}\left(D^{v}\right) \\
&= \operatorname{Ent}(D)-\sum_{v=1}^{V} \frac{\left|D^{v}\right|}{|D|}\left(-\sum_{k=1}^{|\mathcal{Y}|} p_{k} \log_{2} p_{k}\right) \\
&= \operatorname{Ent}(D)-\sum_{v=1}^{V} \frac{\left|D^{v}\right|}{|D|}\left(-\sum_{k=1}^{|\mathcal{Y}|} \frac{\left|D_{k}^{v}\right|}{\left|D^{v}\right|} \log_{2} \frac{\left|D_{k}^{v}\right|}{\left|D^{v}\right|}\right)
\end{aligned}
$$

其中，D_k^v 样本集合D中在属性a上取值为a^v且类别为k的样本

C4.5决策树

C4.5 algorithm

C4.5决策树——以信息增益率为准则来选择划分属性的决策树

信息增益率：

$$\operatorname{Gain\_ratio}(D, a)=\frac{\operatorname{Gain}(D, a)}{\operatorname{IV}(a)}$$

其中

$$\operatorname{IV}(a)=-\sum_{v=1}^{V} \frac{\left|D^{v}\right|}{|D|} \log_{2} \frac{\left|D^{v}\right|}{|D|}$$

CART决策树

CART algorithm

CART决策树——以基尼指数为准则来选择划分属性的决策树

基尼值： $$\operatorname{Gini}(D)=\sum_{k=1}^{|\mathcal{Y}|}\sum_{k'\neq k}p_k p_{k'}=\sum_{k=1}^{|\mathcal{Y}|}p_k\sum_{k'\neq k}p_{k'}=\sum_{k=1}^{|\mathcal{Y}|}p_k(1-p_k)=1-\sum_{k=1}^{|\mathcal{Y}|}p_k^2$$

基尼指数： $$\operatorname{Gini\_index}(D,a)=\sum_{v=1}^{V}\frac{|D^v|}{|D|}\operatorname{Gini}\left(D^v\right)$$

基尼值和基尼指数越小，样本集合纯度越高。

CART决策树

CART algorithm

CART决策树分类算法：

1.根据基尼指数公式 $\text{Gini\_index}(D,a)=\sum_{v=1}^{V}\frac{|D^v|}{|D|}\text{Gini}(D^v)$ 找出基尼指数最小的属性 a_*

2.计算属性 a_* 的所有可能取值的基尼值 $\text{Gini}(D^v), v=1,2,..,V$，选择基尼值最小的取值 a_*^v 作为划分点，将集合D划分为D1和D2两个集合（节点），其中D1集合的样本为 $a_*=a_*^v$ 的样本，D2集合为 $a_*\neq a_*^v$ 的样本

3.对集合D1和D2重复步骤1和步骤2，直至满足停止条件

CART决策树

CART algorithm

CART决策树回归算法:

1.根据以下公式找出最优划分特征 a_* 和最优划分点 a_*^v :

$$a_*, a_*^v = \arg\min_{a,a^v}\left[\min_{c_1}\sum_{\boldsymbol{x}_i\in D_1(a,a^v)}(y_i-c_1)^2+\min_{c_2}\sum_{\boldsymbol{x}_i\in D_2(a,a^v)}(y_i-c_2)^2\right]$$

其中, $D_1(a,a^v)$ 表示在属性a上取值小于等于a^v 的样本集合, $D_2(a,a^v)$ 表示在属性a上取值大于 a^v 的样本集合, c_1表示 D_1 的样本输出均值, c_2 表示 D_2 的样本输出均值

2.根据划分点 a_*^v 将集合D划分为 D_1 和 D_2 两个集合（节点）

3.对集合 D_1 和 D_2 重复步骤1和步骤2, 直至满足停止条件

结 语

在这次课程中，我们学习了西瓜书

决策树的公式推导

那么在下次课程中，我们将会学习西瓜书

支持向量机的公式推导

deepshare.net
深度之眼

联系我们:

电话：18001992849

邮箱：service@deepshare.net

Q Q： 2677693114

公众号

客服微信