1. 决策树

信息熵 $Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$

$D=\{(x_1,y_1),(x_2,y_2)\cdots(x_m,y_m)\}$ 表示样本集合，$|y|$ 表示样本类别总数，2分类为2，多分类为n。p_k 表示第k类样本所占的比例，且 $0\le p_k\le 1$，$\sum_{k=1}^{|y|} p_k = 1$。

$Ent(D)$ 值越小，纯度越高。

证明：　$0\le Ent(D)\le \log_2|y|$

求 $Ent(D)$ 最大值：

若令 $|y|=n$，$p_k = x_k$，那么信息熵 $Ent(D)$ 就可以看作一个n元实值函数，也即 $Ent(D)=f(x_1,x_2,\cdots x_n) = -\sum_{k=1}^{n} x_k \log_2 x_k$，其中 $0\le x_k\le 1$，$\sum_{k=1}^{n} x_k = 1$。

下面考虑多元函数最值。

如果不考虑约束 $0\le x_k\le 1$，仅考虑 $\sum_{k=1}^{n} x_k=1$ 的话，对 $f(x_1,\cdots x_n)$ 求最大值等价于如下

$$\min \sum_{k=1}^{n} x_k \log_2 x_k$$
$$s.t. \sum_{k=1}^{n} x_k = 1$$

显然，在 $0\le x_k\le 1$ 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。

根据拉格朗日乘子法可知，该优化问题的拉格朗日函数为

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

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

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

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

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

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

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

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

∴ 可以解得 $x_1=x_2=\cdots x_n=\frac{1}{n}$

又∵ 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 $x_1=x_2=\cdots=x_n=\frac{1}{n}$ 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 $f(x_1,\cdots x_n)$ 的最大值点，将 $x_1=x_2=\cdots=x_n=\frac{1}{n}$ 代入 $f(x_1,\cdots x_n)$ 中可得

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

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

求 $E_n(t_n)$ 最小值：

如果不考虑约束 $\sum_{k=1}^{n} x_k=1$，仅考虑 $0\le x_k\le 1$ 的话，$f(x_1,\cdots x_n)$ 可以看做是 n 个互不相关的一元函数的加和，也即 $f(x_1,\cdots x_n)=\sum_{k=1}^{n} g(x_k)$ 其中，$g(x_k)=-x_k\log_2 x_k$，$0\le x_k\le 1$. 那么当 $g(x_1),g(x_2),\cdots g(x_n)$ 分别取到其最小值时 $f(x_1\cdots x_n)$ 也就取到了最小值. 由于 $g(x_1),g(x_2),\cdots g(x_n)$ 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 $g(x_1),\cdots g(x_n)$ 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.

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

$g''(x_1) = \frac{d(g'(x_1))}{dx_1} = \frac{d(-\log_2 x_1 - \frac{1}{\ln 2})}{dx_1} = -\frac{1}{x_1 \ln 2} < 0 \quad 0 \le x_1 \le 1$

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

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

∴ $g(x_1)$最小值为0，同理可得$g(x_2), \dots g(x_n)$的最小值也为0，但是，此时是只考虑$0 \le x_k \le 1$时取得的最小值，若考虑约束$\sum x_k = 1$的话，那么$f(x_1, \dots x_n)$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$\sum_{k=1}^{n} x_k = 1$，可知$x_1 = x_2 = \dots x_{k-1} = x_{k+1} = \dots = x_n = 0$，将其代入$f(x_1, \dots x_n)$可得

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

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

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

$= Ent(D) - H(D|a)$ 信息熵－条件熵

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.

$= Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \left(-\sum_{k=1}^{|y|} p_k \log_2 p_k \right)$

$= Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \left(-\sum_{k=1}^{|y|} \frac{|D^v_k|}{|D^v|} \log_2 \frac{|D^v_k|}{|D^v|} \right)$

其中D^v_k样本集合D中在属性a上取值为a^v且类别为k的样本

4. 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：$Gain\_ratio(D,a) = \frac{Gain(D,a)}{IV(a)}$

其中 $IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$

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：$Gini(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'}$

$= \sum_{k=1}^{|y|} p_k(1-p_k) = 1 - \sum_{k=1}^{|y|} p_k^2$

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

CART决策树分类算法：

1.根据基尼指数公式 $Gini\_index(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v)$ 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 $Gini(D^v)$, $v=1,2,\dots V$. 选择基尼值最小的取值 a_*^v 作为划分点，将集合D划分为 D_1 和 D_2 两个集合(分支)，其中 D_1 集合的样本为 $a_* = a_*^v$ 的样本，D_2 集合为 $a_* \neq a_*^v$ 的样本.

3.对集合 D_1 和 D_2 重复步骤1和步骤2，直到满足停止条件.

5.

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：

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

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。