机器学习总结

机器学习常用算法

学　　院：数学学院

专　　业：统计学

学生姓名：OscarLi

2019 年 12 月 14 日

本作品采用知识共享署名-非商业性使用 4.0 国际许可协议进行许可。访问 http://creativecommons.org/licenses/by-nc/4.0/查看该许可协议。

目录

CHAPTER

1 FP-GROW 树

1.1 介绍

FP-growth(Frequent Pattern Tree, 频繁模式树), 是韩家炜老师提出的挖掘频繁项集的方法，是将数据集存储在一个特定的称作 FP 树的结构之后发现频繁项集或频繁项对，即常在一块出现的元素项的集合 FP 树。FP-growth 算法比 Apriori 算法效率更高，在整个算法执行过程中，只需遍历数据集 2 次，就能够完成频繁模式发现，其发现频繁项集的基本过程如下：（1）构建 FP 树（2）从 FP 树中挖掘频繁项集

FP-growth 的一般流程如下：1：先扫描一遍数据集，得到频繁项为 1 的项目集，定义最小支持度（项目出现最少次数），删除那些小于最小支持度的项目，然后将原始数据集中的条目按项目集中降序进行排列。2：第二次扫描，创建项头表（从上往下降序），以及 FP 树。3：对于每个项目（可以按照从下往上的顺序）找到其条件模式基（CPB，conditional patten base），递归调用树结构，删除小于最小支持度的项。如果最终呈现单一路径的树结构，则直接列举所有组合；非单一路径的则继续调用树结构，直到形成单一路径即可。

1.2 内容

1.2.1 构建 FP 树

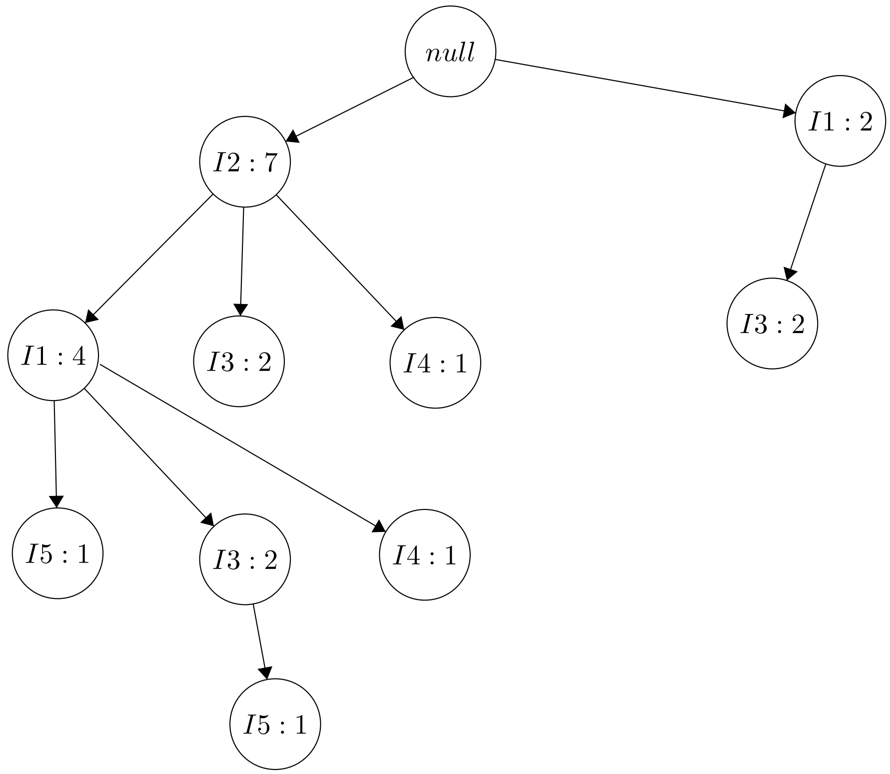
Tid	Items
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I1, I2, I3, I5
9	I1, I2, I3

扫描数据集，对每个物品进行计数：

项集	支持度
I2	7
I1	6
I3	6
I4	2
I5	2

设最小支持度计数为 2 扫描数据库，统计支持度计数，得到频繁 1-项集，按支持度降序排列将其重新排列

Tid	Items
1	I2,I1,I5
2	I2, I4
3	I2, I3
4	I2,I1,I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I2,I1,I3,I5
9	I2,I1, I3



1.2.2 挖掘 FP 树

得到了 FP 树和项头表以及节点链表，我们首先要从项头表的底部项依次向上挖掘。对于项头表对应于 FP 树的每一项，我们要找到它的条件模式基。所谓条件模式基是以我们要挖掘的节点作为叶子节点所对应的 FP 子树。得到这个 FP 子树，我们将子树中每个节点的的计数设置为叶子节点的计数，并删除计数低于支持度的节点。从这个条件模式基，我们就可以递归挖掘得到频繁项集了。

从 I5 开始，对于头表中的每个 I_i，确定自身为频繁模式，再挖掘以 I_i 为后缀的频繁模式

将所有的祖先节点计数设置为叶子节点的计数前缀路径/条件模式基：

<I2,I1:1>、<I2,I1,I3:1>

频繁模式：$\{I2, I5:2\}$ ($\{I2, I5:2\}$(2 的含义指的是 I2，I5 都是 2)

$\{I1, I5:2\}$

$\{I2, I1, I5:2\}$

提取以 I4 的前缀路径/条件模式基：<I2,I1:1>、<I2:1>(I1 被砍掉了) I4 为后缀的频繁模式:$\{I2, I4:2\}$

提取以 I3 为后缀的频繁模式有两分支，则：对于条件 FP 树头表中的每个 Ii，与 I3 连接确定频繁模式 Ii, I3，支持度等于 Ii 的支持度。递归挖掘条件 FP 树，提取以 Ii, I3 为后缀的频繁模式频繁模式：$\{I1, I3:4\}$ $\{I2, I3:4\}$ $\{I2, I1, I3:2\}$

以 I1 为后缀 $\{I2I1:4\}$

I2 只有前缀，没有后缀

项	条件模式基	条件 FP 树	产生的频繁模式
I5	{I2 I1:1}, {I2 I1 I3:1}	<I2:2,I1:2>	{I2 I5:2}, {I1 I5:2}, {I2 I1 I5:2}
I4	{I2 I1:1}, {I2:1}	<I2:2>	{I2 I4:2}
I3	{I2 I1:2}, {I2:2}, {I1:2}	<I2:4, I1:2>, <I1:2>	{I2 I3:4}, {I1 I3:4}, {I2 I1 I3:2}
I1	{I2:4}	<I2:4>	{I2 I1:4}

CHAPTER

2 朴素贝叶斯

2.1 算法

假设有 n 个类别 C1,C2...Cn，给定一个实例的特征向量 w，则此实例属于类 Ci 的概率为

$$P\left(C_i|w\right)=\frac{P\left(w|C_i\right)P\left(C_i\right)}{P(w)}$$

P(Ci) 的计算：将训练样本中属于类 Ci 的实例数量除以训练样本数量即 P(Ci)，例如动物图片识别中，假设有 100 个训练实例，其中有 15 张为猫，则 P(猫) = 15 / 100 = 0.15
P(w) 的计算：因为利用贝叶斯进行分类时，我们只要比较概率的大小即可，而 P(w) 对于所有的类别都是一样的，因此无须计算
P(w|Ci) 的计算：

$$\mathrm{P}\left(\mathrm{w}_0,\mathrm{w}_1,\mathrm{w}_2\cdots\mathrm{w}_\mathrm{n}|\mathrm{C}_\mathrm{i}\right)$$

朴素贝叶斯假设实例的各个属性互相独立,互不影响,因此,上式等价于:$\mathrm{P}\left(\mathrm{w}_0|\mathrm{C}_\mathrm{i}\right)\mathrm{P}\left(\mathrm{w}_1|\mathrm{C}_\mathrm{i}\right)\mathrm{P}\left(\mathrm{w}_2|\mathrm{C}_\mathrm{i}\right)\ldots\mathrm{P}\left(\mathrm{w}_\mathrm{n}|\mathrm{C}_\mathrm{i}\right)$

例子 2.1.1 *假设一个实例的特征向量为 (有四条腿, 会飞)，即 w_0= 有四条腿，w_1 为会飞，共有三个类别分别是鸟、狗、鱼，*
则 $P(w_0|C_0)$=P(有四条腿 / 鸟) = 训练样本中有四条腿的鸟 (实例) 的数量除以样本中鸟 (实例) 的数量
$P(w_1|C_0)$=P(会飞 / 鸟) = 训练样本中会飞的鸟 (实例) 的数量除以样本中鸟 (实例) 的数量

$$P(w_0,w_1|C_0)=P(w_0|C_0)\times P(w1|C0)$$

P(有四条腿, 会飞 / 鸟)= P(有四条腿 / 鸟) P(会飞 / 鸟)*

试由表的训练数据学习一个朴素贝叶斯分类器并确定 $x=(2,S)^T$ 的类标记 y 表中 X(1)，X(2) 为特征，取值的集合分别为 A1 = 1,2,3，A2 = S，M,L，Y 为类标记，Y∈ C = 1,-1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X(1)	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
X(2)	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

$$P(Y=1)=\frac{9}{15}, \quad P(Y=-1)=\frac{6}{15}$$

$$P\left(X^{(1)}=1|Y=1\right)=\frac{2}{9}, \quad P\left(X^{(1)}=2|Y=1\right)=\frac{3}{9}, \quad P\left(X^{(1)}=3|Y=1\right)=\frac{4}{9}$$

$$P\left(X^{(2)}=S|Y=1\right)=\frac{1}{9}, \quad P\left(X^{(2)}=M|Y=1\right)=\frac{4}{9}, \quad P\left(X^{(2)}=L|Y=1\right)=\frac{4}{9}$$

$$P\left(X^{(1)}=1|Y=-1\right)=\frac{3}{6}, \quad P\left(X^{(1)}=2|Y=-1\right)=\frac{2}{6}, \quad P\left(X^{(1)}=3|Y=-1\right)=\frac{1}{6}$$

$$P\left(X^{(2)}=S|Y=-1\right)=\frac{3}{6}, \quad P\left(X^{(2)}=M|Y=-1\right)=\frac{2}{6}, \quad P\left(X^{(2)}=L|Y=-1\right)=\frac{1}{6}$$

$P(Y=1)P\left(X^{(1)}=2|Y=1\right)P\left(X^{(2)}=S|Y=1\right)=\frac{9}{15}\cdot\frac{3}{9}\cdot\frac{1}{9}=\frac{1}{45}$
$P(Y=-1)P\left(X^{(1)}=2|Y=-1\right)P\left(X^{(2)}=S|Y=-1\right)=\frac{6}{15}\cdot\frac{2}{6}\cdot\frac{3}{6}=\frac{1}{15}$ $P(Y=-1)P\left(X^{(1)}=2|Y=-1\right)P\left(X^{(2)}=\right.$
最大

CHAPTER

3 决策树

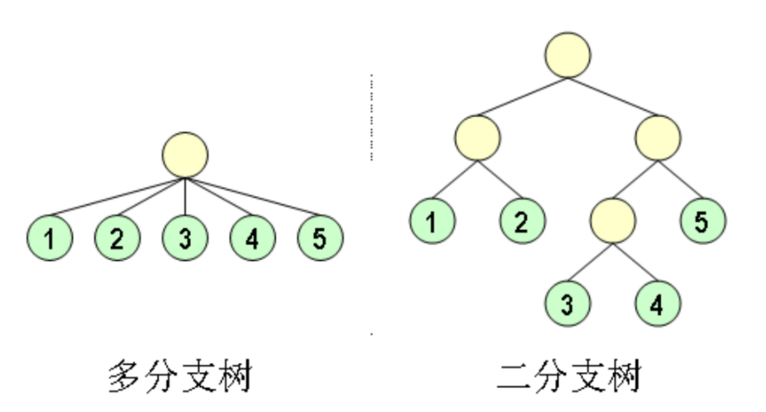


图 3.1: 决策树

设数据集 D 中有 m 个不同的类 C1, C2, C3, ..., Cm 设 Ci,D 是数据集 D 中 Ci 类的样本的集合, |D| 和 |Ci,D| 分别是 D 和 Ci,D 中的样本个数数据集 D 的信息熵:

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

其中 pi 是数据集 D 中任意样本属于类 Ci 的概率

年龄	收入	学生	信用	买了电脑
<30	高	否	一般	否
<30	高	否	好	否
30-40	高	否	一般	是
>40	中等	否	一般	是
>40	低	是	一般	是
>40	低	是	好	否
30-40	低	是	好	是
<30	中	否	一般	否
<30	低	是	一般	是
>40	中	是	一般	是
<30	中	是	好	是
30-40	中	否	好	是
30-40	高	是	一般	是
>40	中	否	好	否

$|D| = 14$
$|C_{1,D}| = 5$
$|C_{2,D}| = 9$

$$Info(D) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14} = 0.940$$

信息增益:

$$\mathrm{Gain}(A) = \mathrm{Info}(D) - \mathrm{Info}_A(D)$$

确定第一次分裂的属性：按年龄划分：
年龄 <30 的有 5 个, 其中 3 个为“否”
年龄 30-40 的有 4 个, 其中 0 个为“否”
年龄 >40 的有 5 个, 其中 2 个为“否”
$Info_{\text{年龄}}D = \frac{5}{14}\left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}\right) + \frac{4}{14}\left(-\frac{4}{4}\log\frac{4}{4} - \frac{0}{4}\log\frac{0}{4}\right) + \frac{5}{14}\left(-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}\right)$
Gain(年龄) = Info(D) - Info 年龄 (D) = 0.940 - 0.694 = 0.246
确定第一次分裂的属性：按收入划分：
收入 = 高的有 4 个, 其中 2 个为“否”
收入 = 中的有 6 个, 其中 2 个为“否”
收入 = 低的有 4 个, 其中 1 个为“否”
$Info_{\text{收入}}D = \frac{4}{14}\left(-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}\right) + \frac{6}{14}\left(-\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6}\right) + \frac{4}{14}\left(-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}\right)$
确定第一次分裂的属性：按信用划分：
信用好的有 6 个, 其中 3 个为“否”
信用一般的有 8 个, 其中 2 个为“否”
$Info_{\text{信用}}D = \frac{6}{14}\left(-\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6}\right) + \frac{8}{14}\left(-\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8}\right)$
确定第一次分裂的属性：按学生划分
是学生的有 7 个, 其中 1 个为“否”
不是学生的有 7 个, 其中 4 个为“否”
$Info_{\text{学生}}D = \frac{7}{14}\left(-\frac{1}{7}\log\frac{1}{7} - \frac{6}{7}\log\frac{6}{7}\right) + \frac{7}{14}\left(-\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7}\right)$

“年龄”属性具体最高信息增益，成为分裂属性

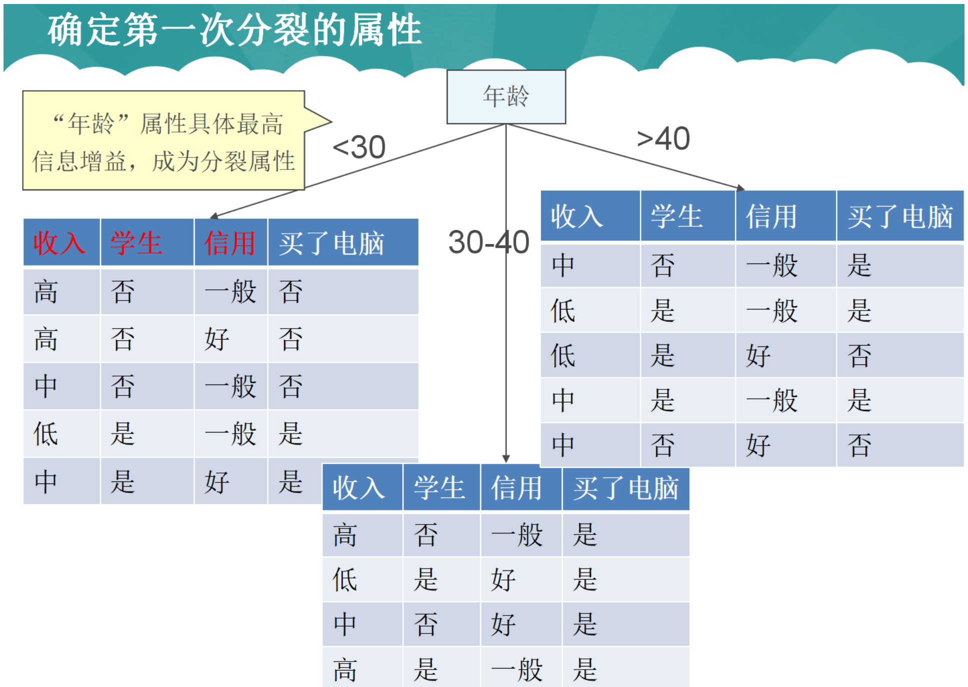


图 3.2: 第二次分类

观察年龄 <30 的人群:

收入	学生	信用	买了电脑
高	否	一般	否
高	否	好	否
中	否	一般	否
低	是	一般	是
中	是	好	是


确定第二次分裂的属性
收入 学生 信用 买了电脑
高 否 一般 否
高 否 好 否
中 否 一般 否
低 是 一般 是
中 是 好 是
“学生”属性具体最高信息增益，成为分裂属性
Info收入(D)
= 2/5 * (-2/2 * log2/2 - 0/2 * log0/2)
+ 2/5 * (-1/2 * log1/2 - 1/2 * log1/2)
+ 1/5 * (-1/1 * log1/1 - 0/1 * log0/1)
= 0.400
Info学生(D)
= 3/5 * (-3/3 * log3/3 - 0/3 * log0/3)
+ 2/5 * (-2/2 * log2/2 - 0/2 * log0/2)
= 0
Info信用(D)
= 3/5 * (-2/3 * log2/3 - 1/3 * log1/3)
+ 2/5 * (-1/2 * log1/2 - 1/2 * log1/2)
= 0.951

图 3.3: 第二次分类

、

CHAPTER

4

BP 神经网络

BP(back propagation) 神经网络是 1986 年由 Rumelhart 和 McClelland 为首的科学家提出的概念，是一种按照误差逆向传播算法训练的多层前馈神经网络，是应用最广泛的神经网络

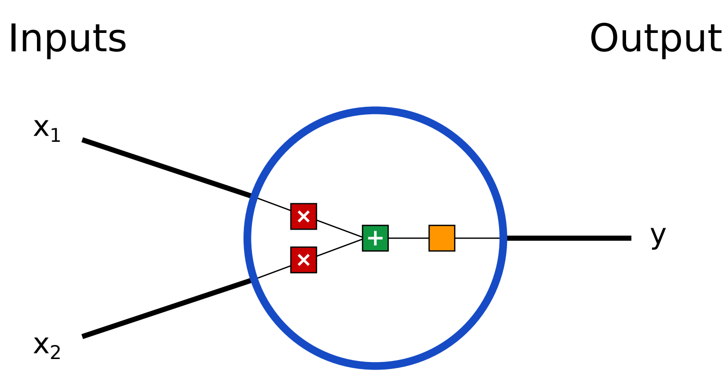


图 4.1: 神经元

注意这里有三件事发生了, 红色代表权重,

$$x_1 \to x_1 * w_1$$
$$x_2 \to x_2 * w_2$$

绿色代表偏差 (bias)

$$(x_1 * w_1) + (x_2 * w_2) + b$$

黄色代表激活函数

$$y = f\left(x_1 * w_1 + x_2 * w_2 + b\right)$$

激活函数常用的是 Sigmod 函数

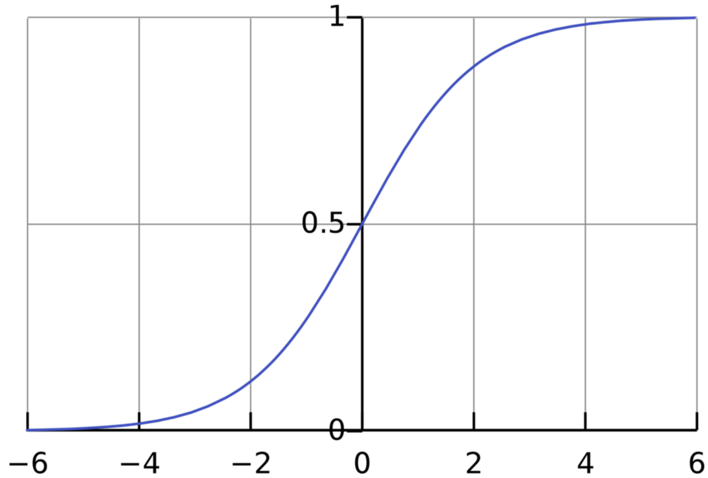


图 4.2: Sigmod

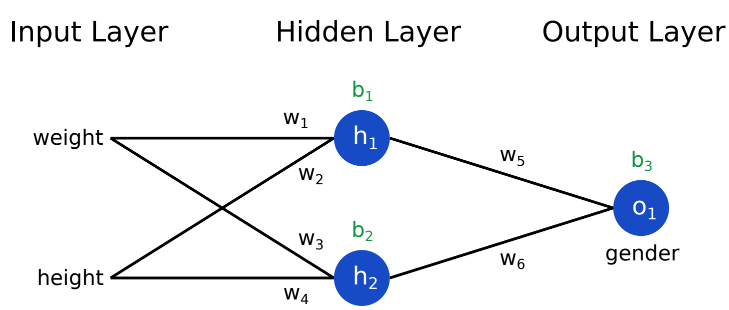


图 4.3: network

$$L\left(w_1, w_2, w_3, w_4, w_5, w_6, b_1, b_2, b_3\right)$$

算法 1 神经网络基本算法

$a_1^{(2)} = g(\Theta_{10}^1 x_0 + \Theta_{11}^1 x_1 + \Theta_{12}^1 x_2 + \Theta_{13}^1 x_3)$
$a_2^{(2)} = g(\Theta_{20}^1 x_0 + \Theta_{21}^1 x_1 + \Theta_{22}^1 x_2 + \Theta_{23}^1 x_3)$
$a_3^{(2)} = g(\Theta_{30}^1 x_0 + \Theta_{31}^1 x_1 + \Theta_{32}^1 x_2 + \Theta_{33}^1 x_3)$

算法 2 反向传播算法

$$
\begin{aligned}
z_i^{(l+1)} &= b_i^{(l)} + \sum_{j=1}^{S_l} w_{ij}^{(l)} a_j^{(l)} \\
g(x) &= \frac{1}{1+e^{-x}} \\
a_i^{(l)} &= g\left(z_i^{(l)}\right) \\
J(\theta) &= \frac{1}{2} \sum_{j=1}^{S_{n_l}} \left(y_j - a_j^{(n_l)}\right)^2 \\
\delta_i^{(l)} &= \frac{\partial J(\theta)}{\partial z_i^{(l)}}
\end{aligned}
$$