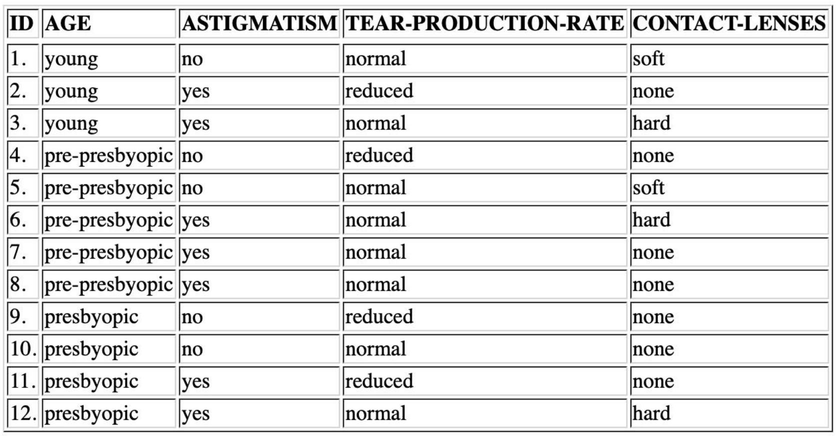
第八讲：监督学习

- [决策树] 基于信息增益，对下述数据集进行决策树构建，描述过程

  一个关于配眼镜的一个决策分类所需要的数据，数据集包含 4 属性：age, astigmatism, trear-prod-rate 为输入特征，contact-lenses 为决策属性。

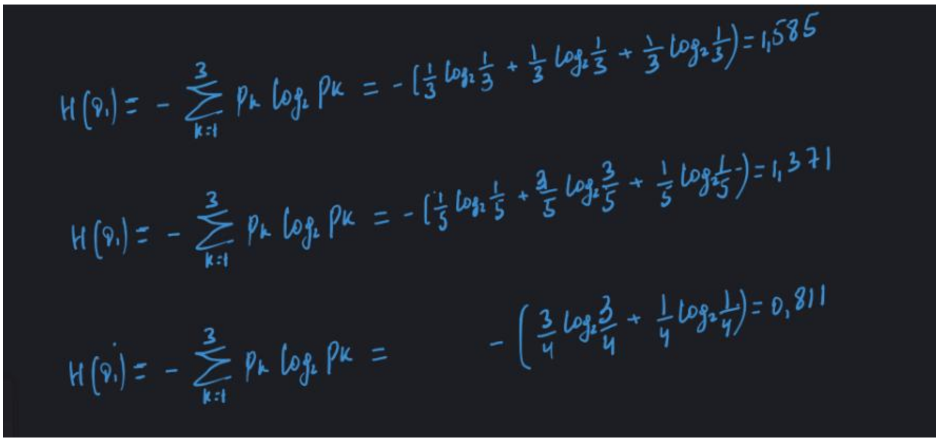
- 

ID	AGE	ASTIGMATISM	TEAR-PRODUCTION-RATE	CONTACT-LENSES
1.	young	no	normal	soft
2.	young	yes	reduced	none
3.	young	yes	normal	hard
4.	pre-presbyopic	no	reduced	none
5.	pre-presbyopic	no	normal	soft
6.	pre-presbyopic	yes	normal	hard
7.	pre-presbyopic	yes	normal	none
8.	pre-presbyopic	yes	normal	none
9.	presbyopic	no	reduced	none
10.	presbyopic	no	normal	none
11.	presbyopic	yes	reduced	none
12.	presbyopic	yes	normal	hard

属性 age 的信息熵 ：

age = young	= {1，2，3}	soft 1/3 none 1/3 hard 1/3
age = pre_presbyopc	= {4，5，6,7,8}	soft 1/5 none 3/5 hard 1/5
age = byopic	= {9，10，11,12}	soft 0/4 none 3/4 hard 1/3

分别：young,pre_presbyopc,byopic 的信息熵为：1.585，1.371，0.811

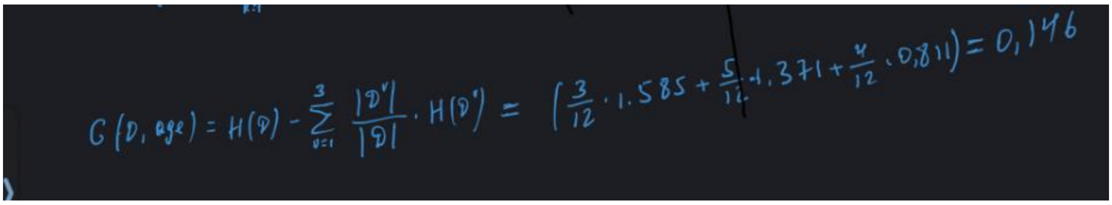


$$H(q_1) = -\sum_{k=1}^{3} p_k \log_2 p_k = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 1{,}585$$

$$H(q_1) = -\sum_{k=1}^{3} p_k \log_2 p_k = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{3}{5}\log_2\frac{3}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 1{,}371$$

$$H(q_1) = -\sum_{k=1}^{3} p_k \log_2 p_k = -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = 0{,}811$$

属性 age 的信息增益为:　0.146



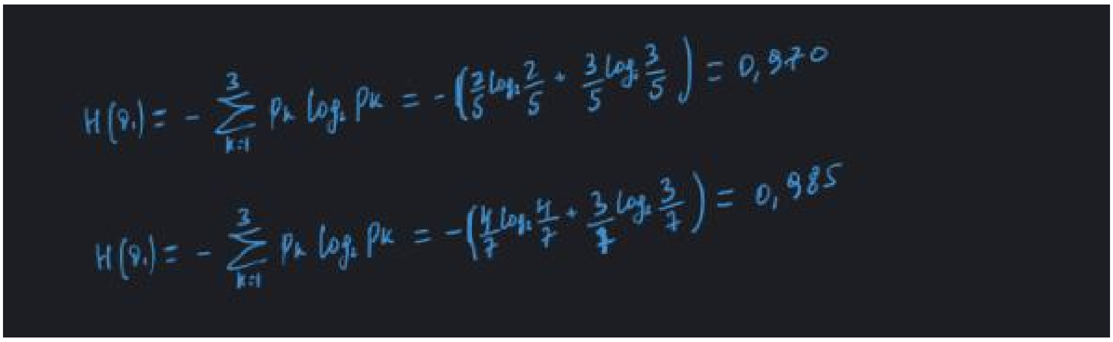
$$G(D, age) = H(D) - \sum_{v=1}^{3} \frac{|D^v|}{|D|} \cdot H(D^v) = \left(\frac{3}{12} \cdot 1.585 + \frac{5}{12} \cdot 1,371 + \frac{4}{12} \cdot 0,811\right) = 0,146$$

属性 astigmatism 的信息熵 ：

Astigmatism no = {1,4,5,9,10}　　　　soft 2/5　none 3/5 hard 0/5

Astigmatism yes = {2,3,6,7,8,11,12}　soft 0/7　none 4/7 hard 3/7

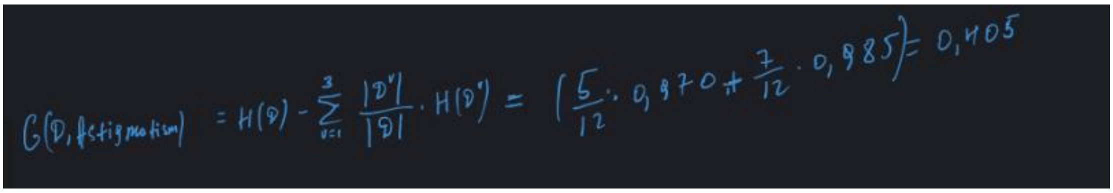
分别：no,yes 的信息熵为：0.970，0.985



$$H(D_1) = -\sum_{k=1}^{3} p_k \log_2 p_k = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0,970$$

$$H(D_1) = -\sum_{k=1}^{3} p_k \log_2 p_k = -\left(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}\right) = 0,985$$

属性 astigmatism 的信息增益为:　0.405



$$G(D, astigmatism) = H(D) - \sum_{v=1}^{3} \frac{|D^v|}{|D|} \cdot H(D^v) = \left(\frac{5}{12} \cdot 0,970 + \frac{7}{12} \cdot 0,985\right) = 0,405$$

属性 tear-production-rate(TPR) 的信息熵 ：

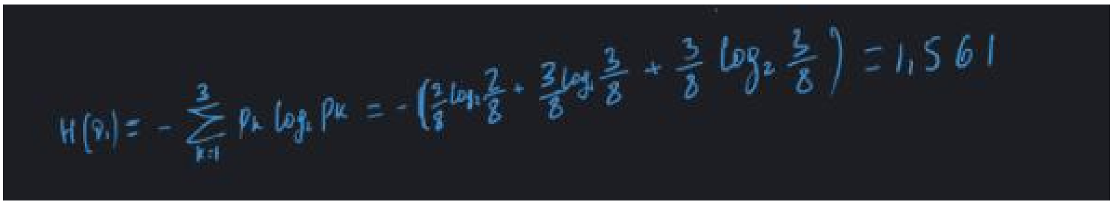
TPR reduced = {2,4,9,11} soft 0/4 none 4/4 hard 0/4

TPR normal = {1,3,5,6,7,8,10,12} soft 2/8 none 3/8 hard 3/8

分别：reduced,normal 的信息熵为： 0，1.561

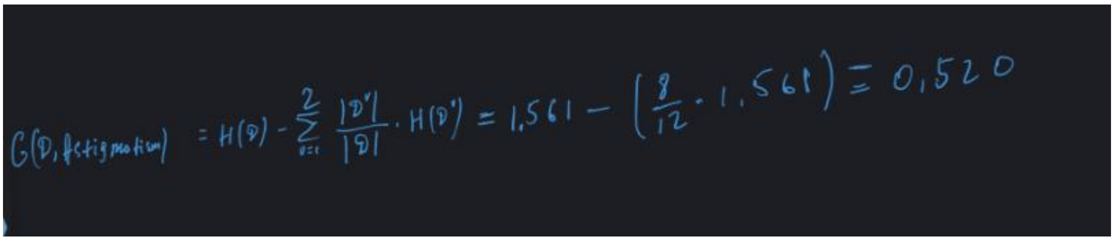
Reduced 因为所有数据集中到一个类所以 H(D_reduced) = 0

H(D_normal) =



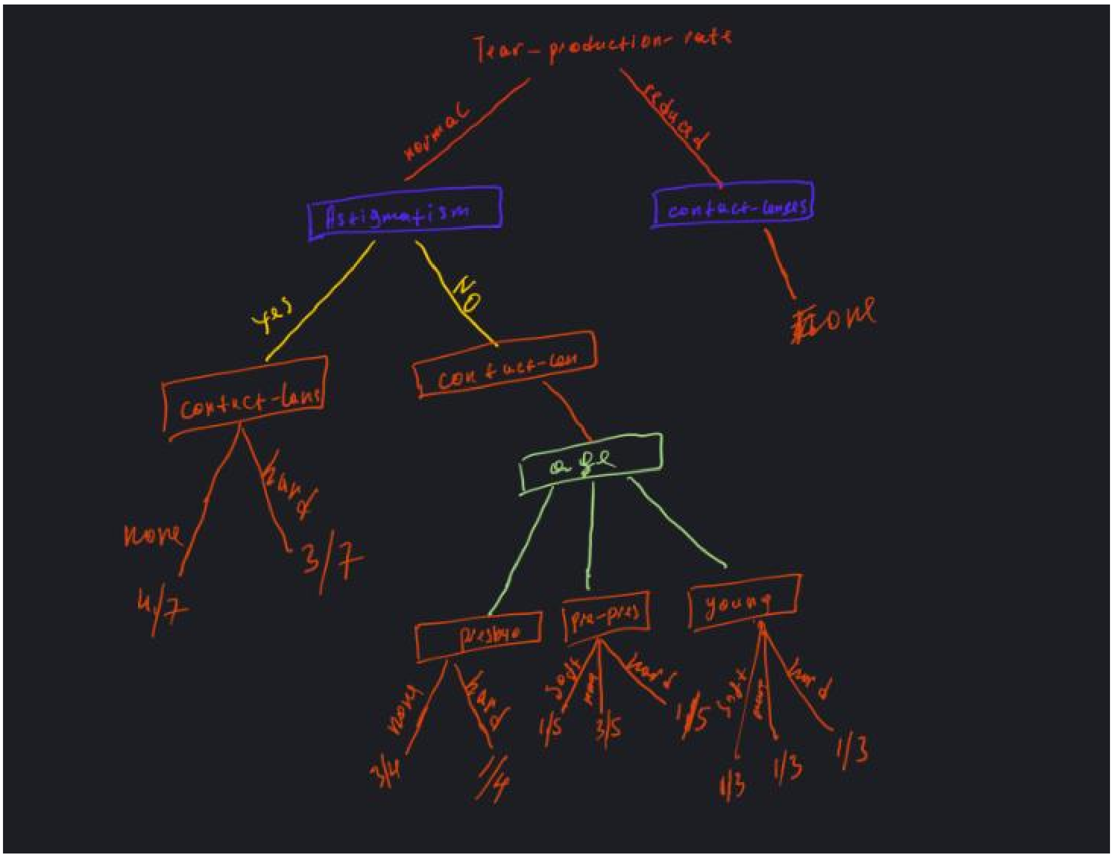
$$H(D_1) = -\sum_{k=1}^{3} p_k \log_2 p_k = -\left(\frac{2}{8}\log_2\frac{2}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{3}{8}\log_2\frac{3}{8}\right) = 1,561$$

属性 TPR 的信息增益为：0.520



$$G(D, \text{Astigmatism}) = H(D) - \sum_{v=1}^{2}\frac{|D^v|}{|D|}\cdot H(D^v) = 1,561 - \left(\frac{8}{12}\cdot 1,561\right) = 0,520$$

Tear-product-rate 的增益最大则把它选为划分属性:


Tear-production-rate
normal
reduced
Astigmatism
contact-lenses
none
yes
NO
contact-lens
contact-len
none
hard
4/7
3/7
age
presbyo
pre-pres
young
none
hard
3/4
1/4
soft
none
hard
1/5
3/5
1/5
soft
none
hard
1/3
1/3
1/3