课后练习4

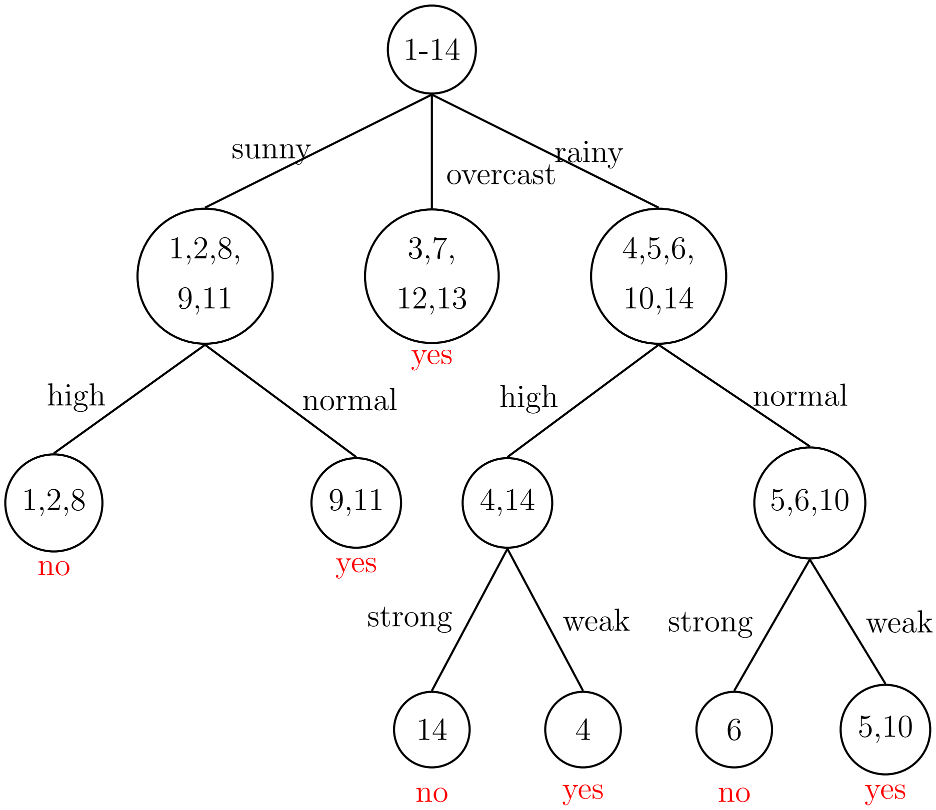
1 问题一

1.1

设四个属性分别为O, T, H, W，则我们可以分别计算它们对应的information gain.

$$
\begin{aligned}
H(D) &= -(\frac{9}{14}\log\frac{9}{14} + \frac{5}{14}\log\frac{5}{14}) \approx 0.94029 \\
g(D,O) &= H(D) - H(D|O) \\
&= H(D) - \big(\frac{5}{14}[-(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5})] + \frac{4}{14}[-(\frac{4}{4}\log\frac{4}{4})] + \frac{5}{14}[-(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5})]\big) \\
&= 0.24675 \\
g(D,T) &= H(D) - H(D|T) \\
&= H(D) - \big(\frac{4}{14}[-(\frac{2}{4}\log\frac{2}{4} + \frac{2}{4}\log\frac{2}{4})] + \frac{6}{14}[-(\frac{4}{6}\log\frac{4}{6} + \frac{2}{6}\log\frac{2}{6})] + \frac{4}{14}[-(\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4})]\big) \\
&= 0.029227 \\
g(D,H) &= H(D) - H(D|H) \\
&= H(D) - \big(\frac{7}{14}[-(\frac{3}{7}\log\frac{3}{7} + \frac{4}{7}\log\frac{4}{7})] + \frac{7}{14}[-(\frac{1}{7}\log\frac{1}{7} + \frac{6}{7}\log\frac{6}{7})]\big) \\
&= 0.15184 \\
g(D,W) &= H(D) - H(D|W) \\
&= H(D) - \big(\frac{8}{14}[-(\frac{2}{8}\log\frac{2}{8} + \frac{6}{8}\log\frac{6}{8})] + \frac{6}{14}[-(\frac{3}{6}\log\frac{3}{6} + \frac{3}{6}\log\frac{3}{6})]\big) \\
&= 0.04813
\end{aligned}
$$

根据计算得到的information gain，我们可以相应建树:



1.2

根据得到的树模型可知，此时到达了{9,11}节点，应该进行户外活动

2 问题二

2.1

使用Bootstrapping方法，从N个样本中每次选一个，有放回地选择$N' = pN$次，那么一个样本未被选中的概率即为

$$\begin{aligned}P &= (1-\frac{1}{N})^{N'}\\ &= (1-\frac{1}{N})^{pN}\end{aligned}$$

当N趋于无穷时，

$$\begin{aligned}&\lim_{N\to\infty}(1-\frac{1}{N})^{pN}\\ =&\big(\lim_{N\to\infty}(1-\frac{1}{N})^{N}\big)^{p}\\ =&(\frac{1}{e})^{p}\\ =&e^{-p}\end{aligned}$$

因此总共有$N \cdot e^{(-p)}$个样本不会被采样到

2.2

一个样本被分类错误，当且仅当有至少两个决策树的分类是错误的. 因此被错误分类的总概率为

$$
\begin{aligned}
E(G) &= E(g_1)E(g_2)E(g_3) + E(g_1)E(g_2)(1 - E(g_3)) + E(g_1)(1 - E(g_2))E(g_3) + (1 - E(g_1))E(g_2)E(g_3) \\
&= 0.15 \times 0.25 \times 0.3 + 0.15 \times 0.25 \times 0.7 + 0.15 \times 0.75 \times 0.3 + 0.85 \times 0.25 \times 0.3 \\
&= 0.135
\end{aligned}
$$

3 问题三

3.1

已知平方损失函数为

$$
L(h) = \sum_{(x_i, y_i) \in D} (y_i - g^{(T)}(x_i))^2
$$

设$g^{(t-1)}(x_i)$与y_i的残差为$r_i^{(t)}$:

$$
r_i^{(t)} = y_i - g^{(t-1)}(x_i)
$$

求偏导得到

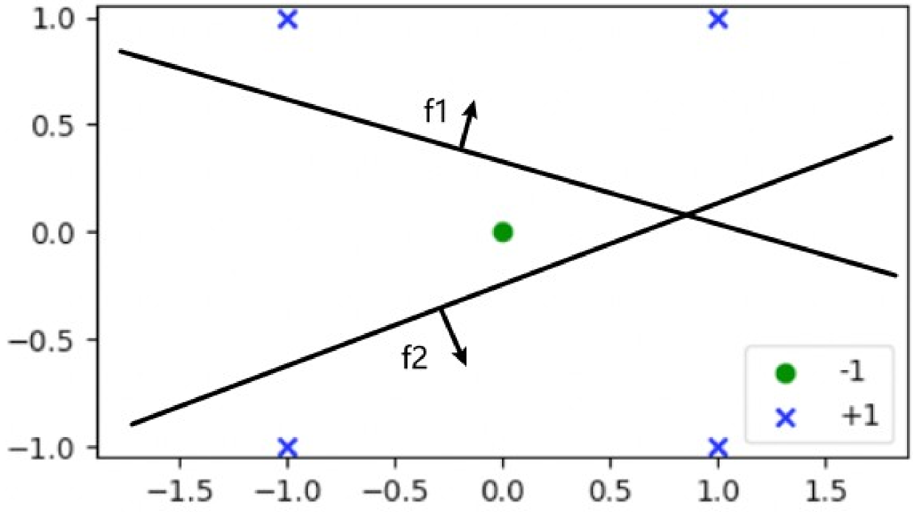
$$
\begin{aligned}
\frac{\partial L}{\partial \alpha_t} &= \sum_{(x_i, y_i) \in D} 2(y_i - g^{(T-1)}(x_i) - \alpha_t f_t(x_i))(-f_t(x_i)) = 0 \\
&\Leftrightarrow \sum_{(x_i, y_i) \in D} (y_i - g^{(T-1)}(x_i) - \alpha_t f_t(x_i)) f_t(x_i) = 0 \\
&\Leftrightarrow \sum_{(x_i, y_i) \in D} f_t(x_i)(y_i - g^{(T-1)}(x_i)) = \alpha_t \sum_{(x_i, y_i) \in D} f_t^2(x_i) \quad (f_t^2(x_i) = 1) \\
&\Leftrightarrow \quad \alpha_t = \frac{\sum_{(x_i, y_i) \in D} f_t(x_i) r_i^T}{|D|}
\end{aligned}
$$

3.2

T最小为2，如图所示，$\alpha_1 = \alpha_2 = 1$

3.3

可以从以下几个角度来缓解Adaboost过拟合:



1. 增大数据集

2. 降低弱分类器f_t的复杂度

3. 增加正则化项，使用L1或L2正则化

4 问题四

4.1

协方差矩阵为

$$K = \begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & k(x_1,x_3) \\ k(x_2,x_1) & k(x_2,x_2) & k(x_2,x_3) \\ k(x_3,x_1) & k(x_3,x_2) & k(x_3,x_3) \end{pmatrix} = \begin{pmatrix} 1 & 0.61 & 0.14 \\ 0.61 & 1 & 0.61 \\ 0.14 & 0.61 & 1 \end{pmatrix}$$

4.2

$$\begin{aligned} \mu(x^*) &= k^T(x^*)K^{-1}y \\ &= \begin{pmatrix} k(x_1,x_4) & k(x_2,x_4) & k(x_3,x_4) \end{pmatrix} K^{-1} \begin{pmatrix} 0.6 \\ 1.5 \\ 3.2 \end{pmatrix} \\ &= 2.23 \\ \sigma^2(x^*) &= k(x^*,x^*) - k^T(x^*)K^{-1}k(x^*) \\ &= 1 - k^T(x^*)K^{-1}k(x^*) \\ &= 0.52 \end{aligned}$$

4.3

$$
\begin{aligned}
\mu(x^*) &= k^T(x^*)(K+\sigma^2 I)^{-1}y \\
&= \begin{pmatrix} k(x_1,x_4) & k(x_2,x_4) & k(x_3,x_4) \end{pmatrix} \begin{pmatrix} 1.5 & 0.61 & 0.14 \\ 0.61 & 1.5 & 0.61 \\ 0.14 & 0.61 & 1.5 \end{pmatrix}^{-1} \begin{pmatrix} 0.6 \\ 1.5 \\ 3.2 \end{pmatrix} \\
&= 1.28 \\
\sigma^2(x^*) &= k(x^*, x^*) - k^T(x^*)(K+\sigma^2 I)^{-1}k(x^*) \\
&= 1 - k^T(x^*)(K+\sigma^2 I)^{-1}k(x^*) \\
&= 0.74
\end{aligned}
$$