hw4-ans

一、

（1）

label为yes的概率：$P(y=yes)=\frac{9}{14}$

label为no的概率：$P(y=no)=\frac{5}{14}$

整个数据集的熵：$H(C)=-\frac{5}{14}\log_2\frac{5}{14}-\frac{9}{14}\log_2\frac{9}{14}\approx 0.940$

分别计算条件熵：$$\begin{aligned}H(C|\text{outlook})&=\frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}\right)+\frac{4}{14}\times 0+\frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}\right)\\&\approx 0.694\end{aligned}$$

$g(C,\text{outlook})=0.94-0.694=0.246$

$$\begin{aligned}H(C|\text{temperature})=&\frac{4}{14}\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)+\frac{6}{14}\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)\\&+\frac{4}{14}\left(-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4}\right)\\&\approx 0.911\end{aligned}$$

$g(C,\text{temperature})=0.94-0.911=0.0290$

$$\begin{aligned}H(C|\text{humidity})&=\frac{1}{2}\left(-\frac{3}{7}\log_2\frac{3}{7}-\frac{4}{7}\log_2\frac{4}{7}\right)+\frac{1}{2}\left(-\frac{6}{7}\log_2\frac{6}{7}-\frac{1}{7}\log_2\frac{1}{7}\right)\\&\approx 0.788\end{aligned}$$

$g(C,\text{humidity})=0.94-0.788=0.152$

$$\begin{aligned}H(C|\text{wind})&=\frac{8}{14}\left(-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4}\right)+\frac{6}{14}\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)\\&\approx 0.892\end{aligned}$$

$g(C,\text{wind})=0.94-0.892=0.048$

由上可得，outlook带来的信息增益最高，因此选择outlook作为第一步的划分依据。

根据outlook分为可三个子数据集：sunny、overcast 、rainy。

1. 由outlook=overcast中play的取值均为yes，因此该子节点为叶子节点且预测为yes。
2. 我们对outlook取sunny的子数据集进一步进行分割，计算三个特征的条件熵：

   $H(C_{\text{sunny}})=0.971$

   $H(C_{\text{sunny}}|\text{temperature})=0.4$

   $H(C_{\text{sunny}}|\text{humidity})=0$

   $H(C_{\text{sunny}}|\text{wind})=0.951$

   因此选择humidity作为分割特征，进一步分为high和normal两个子数据集，label分别为no和yes。
3. 我们对outlook取rainy的子数据集进一步进行分割，计算三个特征的条件熵得到：

   $H(C_{\text{rainy}})=0.971$

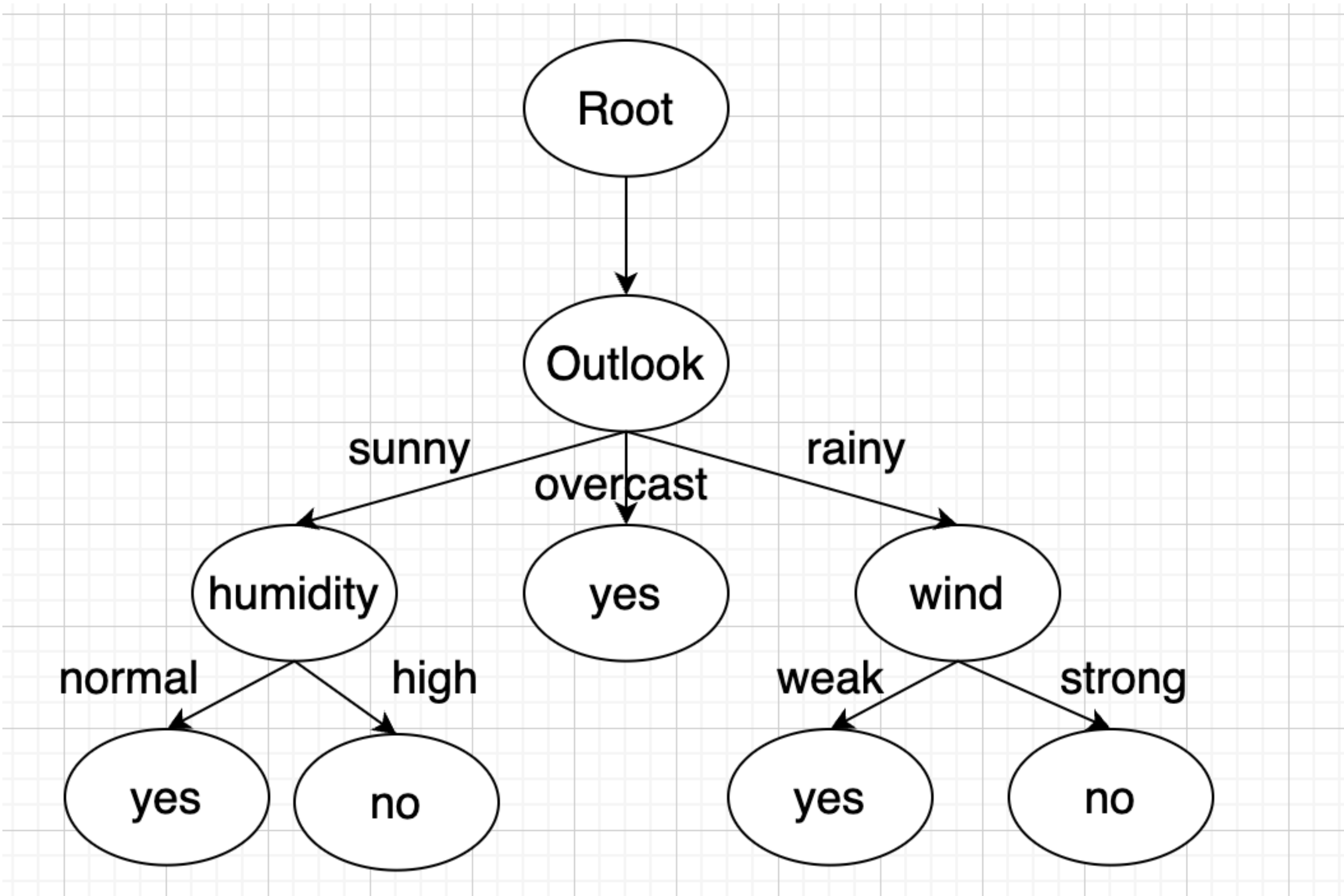
   $H(C_{\text{rainy}}|\text{temperature})=0.951$

   $H(C_{\text{sunny}}|\text{humidity})=0.951$

   $H(C_{\text{sunny}}|\text{wind})=0$

   因此选择wind作为分割特征，进一步分为weak和strong两个子数据集，label分别为yes和no。

因此决策树可构建为：



（2）根据该决策树预测，在上述条件下应当进行户外活动。

二、

（1）自助采样法（bootstrapping）指从原始样本中有放回地采样。每次抽取有N个样本可供选择，第i个样本被采样的概率是$P(已采样)=1/N$。因此，未被采样的概率为$P(未采样)=1-1/N$。那么N'次中，样本未被抽取的概率可以表示为：

$$P=(1-\frac{1}{N})^{N'}=(1-\frac{1}{N})^{pN} \tag{1}$$

当$N\rightarrow\infty$时，可得$P\rightarrow e^{-p}$，因此不会被抽入样本数目为Ne^{-p}。

（2）由二分类问题最终的类别由多数树的输出决定可知，当至少有两棵树判定错误时，随机森林的判定结果错误。由此，

$$\begin{aligned} E(G)&=E(g_1)*E(g_2)*E(g_3)+(1-E(g_1))*E(g_2)*E(g_3)+E(g_1)*(1-E(g_2))*E(g_3)+E(g_1)*E(g_2)*(1-E(g_3)) \\ &=0.15*0.25*0.30+0.85*0.25*0.30+0.15*0.75*0.30+0.15*0.25*0.70 \\ &=0.135 \end{aligned} \tag{2}$$

三、

（1）在squared loss下，t时刻时，

$$\begin{aligned} L^{(t)}&=\sum_{(x_i,y_i)\in\mathcal{D}}(y_i-g^{(t)}(x_i))^2 & (3)\\ &=\sum_{(x_i,y_i)\in\mathcal{D}}(y_i-g^{(t-1)}(x_i)-\alpha_t f^{(t)}(x_i))^2 & (4) \end{aligned}$$

考虑$f^{(t)}$已经完成优化，因此直接计算：

$$\frac{\partial L^{(t)}}{\partial \alpha_t} = -2 \sum_{(x_i,y_i)\in\mathcal{D}} f^{(t)}(x_i)(y_i - g^{(t-1)}(x_i) - \alpha_t f^{(t)}(x_i)) \tag{5}$$

令$\frac{\partial L^{(t)}}{\partial \alpha_t} = 0$，得到$\alpha_t$极值点为：

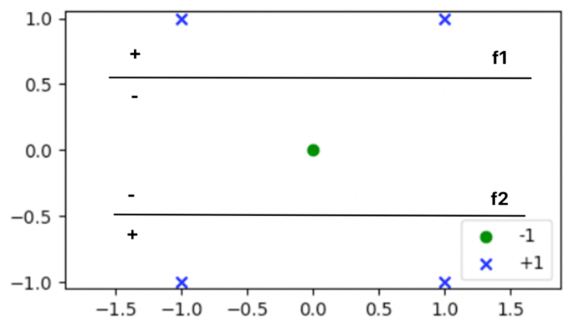
$$\alpha_t = \frac{\sum_{(x_i,y_i)\in\mathcal{D}} f^{(t)}(x_i)(y_i - g^{(t-1)}(x_i))}{\sum_{(x_i,y_i)\in\mathcal{D}} f^{(t)}(x_i)^2} \tag{6}$$

$$= \frac{\sum_{(x_i,y_i)\in\mathcal{D}} f^{(t)}(x_i)(y_i - g^{(t-1)}(x_i))}{N} \tag{7}$$

可验证，此处的α_t为最小值点。

（2）由于该数据集线性不可分，因此$T=1$时不成立。

$T=2$时，考虑$\alpha_0 = \alpha_1 = C$，$(C>0)$，以及如下$f^{(0)}, f^{(1)}$:



容易验证，该实例可以正确分类该数据集。

（3）

- 限制T的最大值，以防止过拟合；
- 设置validation dataset，在其上采用early stopping；
- 将weak classifier f替换成泛化能力更强的模型；
- ....（合理即可）

四、

（1）By definition，协方差矩阵Σ如下：

$$\Sigma = \begin{bmatrix} 1 & e^{-\frac{1}{2}} & e^{-2} \\ e^{-\frac{1}{2}} & 1 & e^{-\frac{1}{2}} \\ e^{-2} & e^{-\frac{1}{2}} & 1 \end{bmatrix} \tag{8}$$

（2）

在无噪声高斯过程中，

$$\begin{gathered} \mu^* = k(x^*)^T K^{-1} y \\ \Sigma^* = k(x^*, x^*) - k(x^*)^T K^{-1} k(x^*) \end{gathered} \tag{9}$$

代入

$$k(\mathbf{x}^*) = \begin{bmatrix} e^{-\frac{9}{2}} \\ e^{-2} \\ e^{-\frac{1}{2}} \end{bmatrix}, \quad k(\mathbf{x}^*, \mathbf{x}^*) = 1 \tag{10}$$

有：

$$\begin{aligned}\mu(\mathbf{x}^*) &\approx 2.22 \\ \sigma^2(\mathbf{x}^*) &\approx 0.52\end{aligned} \tag{11}$$

（3）

在有噪声高斯过程中，

$$\begin{aligned}\mu^* &= k(x^*)^T(K+\sigma^2 I)^{-1}y \\ \Sigma^* &= k(x^*,x^*) - k(x^*)^T(K+\sigma^2 I)^{-1}k(x^*)\end{aligned} \tag{12}$$

代入相关值，有：

$$\begin{aligned}\mu(\mathbf{x}^*) &\approx 1.28 \\ \sigma^2(\mathbf{x}^*) &\approx 0.75\end{aligned} \tag{13}$$