3.17 《机器学习基础》第2次作业

王宇哲 18000111828

T1

证：考虑 $y=c(x)+\varepsilon$，ε为噪声

$\varepsilon \sim \Sigma$，$E[\varepsilon]=0$

则 $E_{T\sim D^m,\varepsilon\sim\Sigma}[(h_T(x)-y)^2]$

$= E_{T,\varepsilon}[h_T^2(x)-2h_T(x)y+y^2]$

$= E_{T,\varepsilon}[h_T^2(x)-2h_T(x)c(x)-2\varepsilon h_T(x)$
$+c^2(x)+2\varepsilon c(x)+\varepsilon^2]$

$= E_T[h_T^2(x)]-2E_T[h_T(x)]c(x)+c^2(x)$
$-2E[\varepsilon]E_T[h_T(x)]+2E[\varepsilon]c(x)+E[\varepsilon^2]$；

（这一步用到ε与$h_T(x)$独立的条件）

考虑 $E_T[h_T^2(x)]-\bar{h}^2(x)=Var(x)$

$\bar{h}^2(x)-2\bar{h}(x)c(x)+c^2(x)=(\bar{h}(x)-c(x))^2$
$=Bias^2(x)$

$E_{T,\varepsilon}[h_T(x)]=\bar{h}(x)$

代入原式，即得

$E_{T\sim D^m,\varepsilon\sim\Sigma}[(h_T(x)-y)^2]$

$= Var(x)+Bias^2(x)+E[\varepsilon^2]$

即为偏差-方差-噪声分解公式 证毕！

T2

解，同：

i) 通过随机采样的方式从数据集D中构建训练集T和测试集T'

ii) 每次随机划分并进行训练时，$T\cap T'=\emptyset$

iii) 都需要重复进行若干次数据集的随机划分过程，以均值作为最终评估结果

异：

i) 交叉验证法通过无放回的抽样方法划分出验证集T'，而自助法通过有放回测试的抽样方法得到T'和T

ii) 交叉验证法训练集T规模小于原始数据集D，而自助法T、D规模相等，即模型选择阶段和最终模型训练阶段数据集规模无差异

iii) 自助法训练集T与原始数据集D中数据分布不一定一致，而（通过分层采样）交叉验证法可保证数据分布大致的一致性；因此自助法不适用对数据分布敏感的模型选择