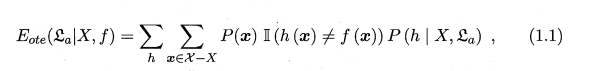
机器学习

西瓜书公式1.1到理解推导

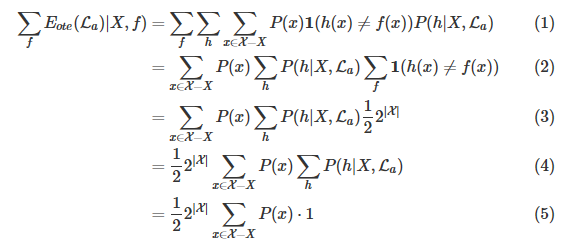
下面按我自己的理解推导一下

原文条件如下：

假设样本空间和H假设空间组都是离散的，令代表算法L*a*基于训练数据X产生假设h的概率，再令f代表我们希望学习的真实目标函数L*a*的"训练集外误差"，即L*a*在训练、集之外的所有样本上的误差为



其中Ⅱ(· )是指示函数，若·为真则取值1，否则取值0



看公式主要是对和的推导理解

拆开来看各个公式的作用：

1. 

是样本，X是训练数据，所以即是训练集外数据，对每个训练集外数据的概率求和，即为训练集外数据的概率。

1. 

指示函数，当假设与真实函数值不等时，代表有误差，该公式表示假设函数对比真实函数在样本上的误差个数，又因为是均匀分布，该公式的值为1/2的样本数量。

当样本有2个属性时，样本空间为=1+2+1

当样本有3个属性时，样本空间为=1+3+3+1

当样本有4个属性时，样本空间为=1+4+6+4+1

当样本有5个属性时，样本空间为=1+5+10+10+5+1

当样本有6个属性时，样本空间为............（末尾1为Ø的情况）

把数字列出来不就是杨辉三角吗

1

1. 1
2. 2 1 2个属性
3. 3 3 1 3个属性
4. 4 6 4 1 4个属性
5. 5 10 10 5 1 5个属性
6. ..................

所有求样本空间实际上就是求杨辉三角一层的和，即（n为层数）

实际属性与层数相差1，即2个属性实际为层和，3个属性时实际为4层和

那么该函数可以简化成，推导出



1. 

已知是代表算法L*a*基于训练数据X产生假设h的概率，该公式表示所有假设的概率和，即为1，推导出



整个公式即为表示在该域所有问题上算法在训练集外上的所有误差。