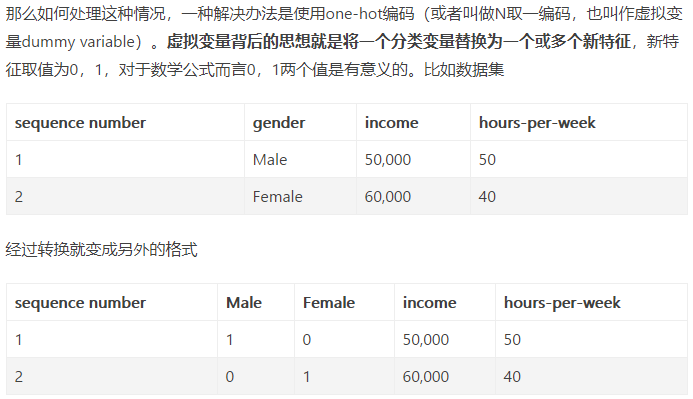
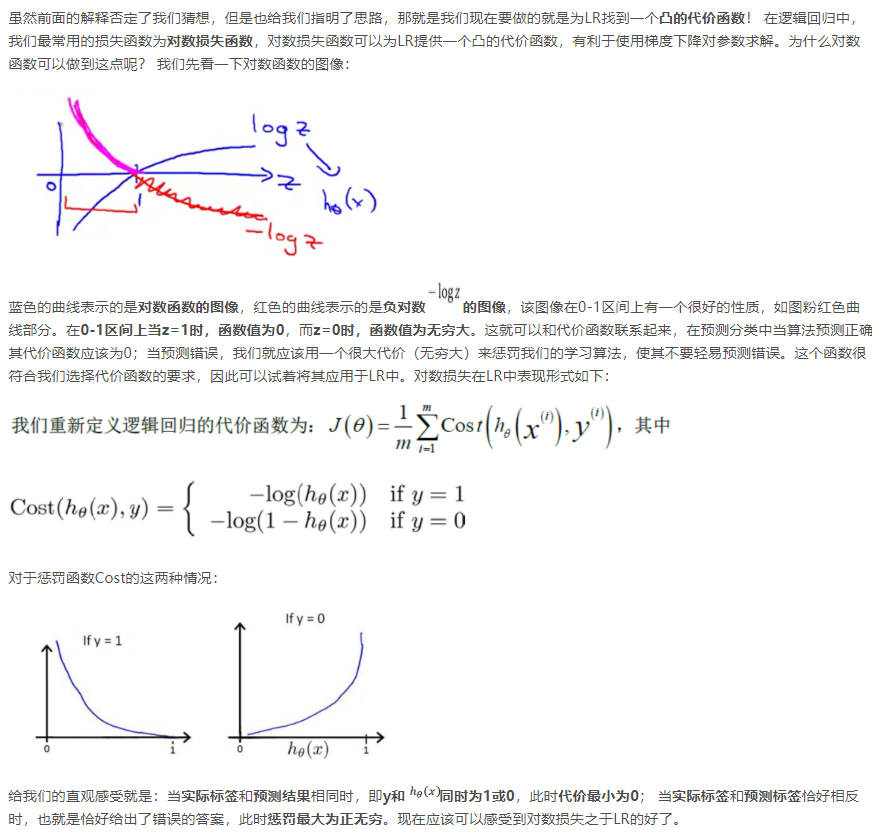
# **建模过程中分类变量的处理**

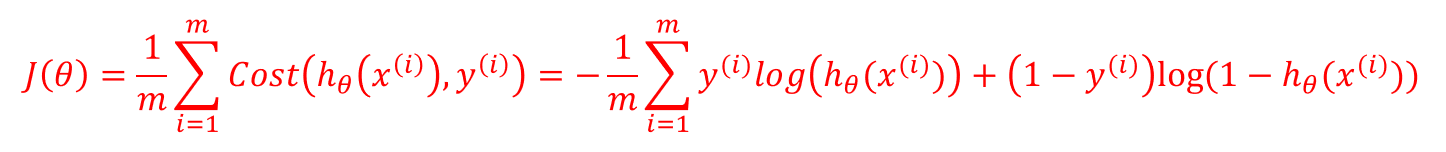


1. 离散特征的取值之间没有大小的意义，比如color：[red,blue],那么就使用one-hot编码
2. 离散特征的取值有大小的意义，比如size:[X,XL,XXL],那么就使用数值的映射{X:1,XL:2,XXL:3}说明：对于有大小意义的离散特征，直接使用映射就可以了，{'XL':3,'L':2,'M':1}

缺失值处理：

逻辑回归：





决策树：

在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。Entropy = 系统的凌乱程度，使用算法[ID3](https://baike.baidu.com/item/ID3" \t "https://baike.baidu.com/item/%E5%86%B3%E7%AD%96%E6%A0%91/_blank), [C4.5](https://baike.baidu.com/item/C4.5" \t "https://baike.baidu.com/item/%E5%86%B3%E7%AD%96%E6%A0%91/_blank)和C5.0生成树算法使用熵。

决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。

建立决策树主要有三种算法

**ID3 :适应信息增益来进行特征选择的决策树学习过程**

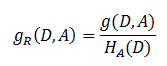
信息增益=信息熵-条件熵。

 缺点：信息增益偏向取值较多的特征

原因：当特征的取值较多时，根据此特征划分更容易得到纯度更高的子集，因此划分之后的熵更低，由于划分前的熵是一定的，因此信息增益更大，因此信息增益比较 偏向取值较多的特征。

IMG_256

C4.5：信息增益率



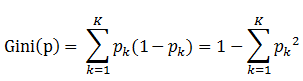
CART：基尼指数

定义：基尼指数（基尼不纯度）：表示在样本集合中一个随机选中的样本被分错的概率。

        注意： Gini指数越小表示集合中被选中的样本被分错的概率越小，也就是说集合的纯度越高，反之，集合越不纯。

即 基尼指数（基尼不纯度）= 样本被选中的概率 \* 样本被分错的概率

书中公式：



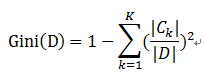
说明:

1. pk表示选中的样本属于k类别的概率，则这个样本被分错的概率是(1-pk)

2. 样本集合中有K个类别，一个随机选中的样本可以属于这k个类别中的任意一个，因而对类别就加和

3. 当为二分类是，Gini(P) = 2p(1-p)

样本集合D的Gini指数 ： 假设集合中有K个类别，则：



总结：一个属性的信息增益越大，表明属性对样本的熵减少的能力越强，这个属性使得数据由不确定性变成确定性的能力越强