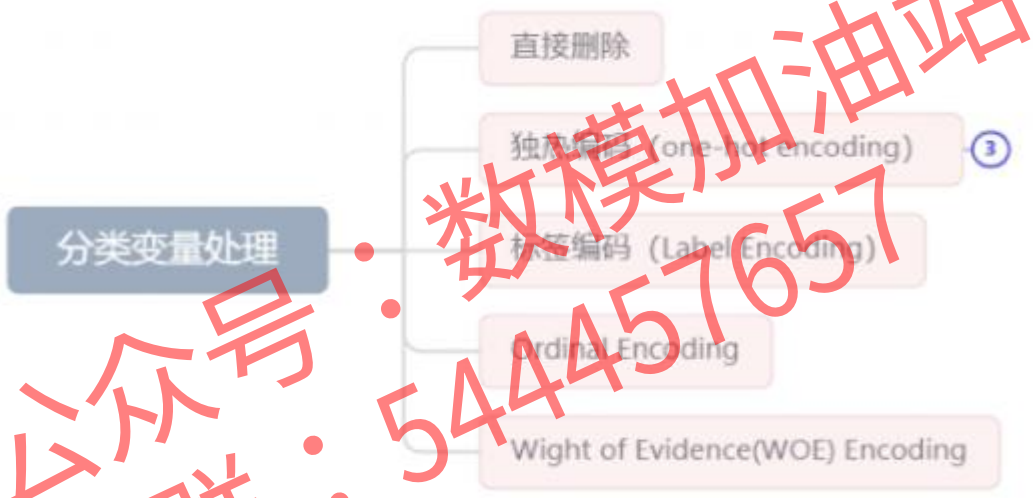
数据变换——分类变量处理

在数据挖掘过程中，算法可以直接处理数值型变量，但是算法一般无法直接处理分类变量。 分类变量没有好坏多少之分，多用 0,1,2,等数值代表一个类型，如果直接引入模型中计算容易让计算机误以为这是一个数值型变量，从而出现错误。

因此，在训练模型之前，需要对分类变量进行处理，使之转换为数值型变量。 常见的分类变量处理方法如下：



这里介绍一下比较热门的独热编码（one-hot Encoding）：

独热编码是一种将分类变量转换为若干二进制列的方法，其中 1 表示属于该类别的样本。

Human-Readable

Pet
Cat
Dog
Turtle
Fish
Cat

Machine-Readable

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

MATLAB 实现方法：

```
function data_encoded = onehot(data)
% 独热编码函数，将分类变量转换若干二进制列
% 输入：分类变量组成的数组 data，注意数组元素均需为整数
% 输出：编码好的数组 data_encoded
  test = abs(rem(data,1));
if sum(sum(test)) ~= 0
    fprintf('数组元素必须全为整数！\n'
   return
 end

[m, n] = size(data);
data_encoded = [];
for k = 1:n
    d = data(:,k);
    labels = unique(d); %d 的类别
    num_labels = length(labels); %类别个数
    data_encoding = zeros(m,num_labels);
    for i = 1:m
        for j = 1:num_labels
            if d(i) == labels(j)
                data_encoding(i,j) = 1;
            else
                continue
            end
        end
    end
    data_encoded =[data_encoded,data_encoding];
end
```

可以在 MATLAB 命令行窗口中输入以下代码试运行 onehot 函数：

```
x = randi(3,5,1)
y = onehot(x)
```

```
x =

   2
   3
   2
   3
   1
```

```
y =

   0   1   0
   0   0   1
   0   1   0
   0   0   1
   1   0   0
```