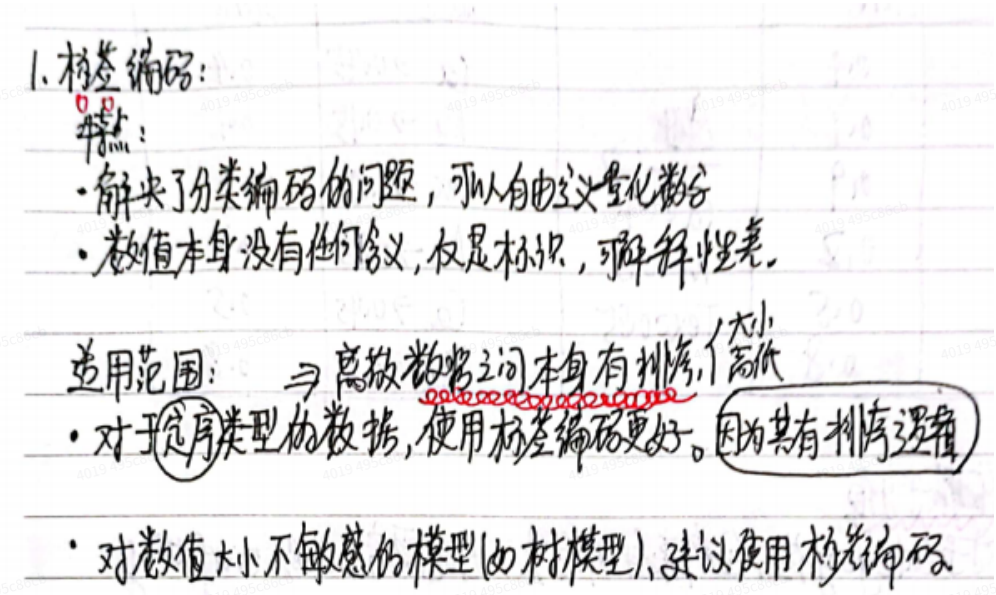
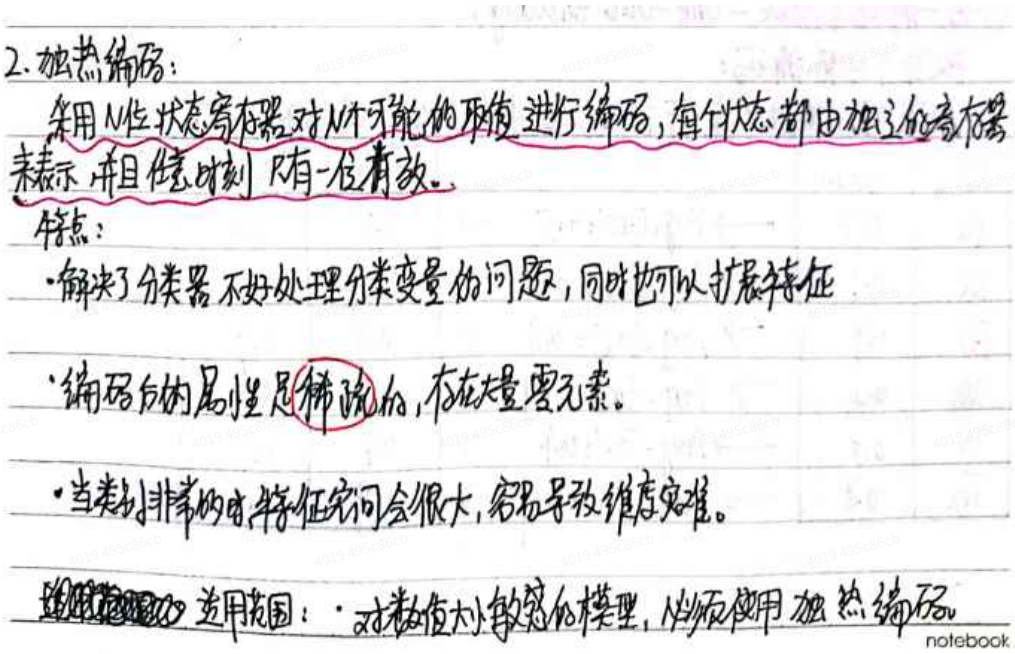
7.编码方式

由于机器学习算法都是在矩阵上执行线性代数计算的，**所以参加计算的特征必须是数值型的。**对于非数值型的特征需要进行编码处理。





Date: \ \ Page:

3) 目标编码 (target encoding)

使用target值的均值作为 category 变量的替换值。

State	Score
Ca	0.4
Ne	0.1
Tex	0.9
Ne	0.2
Ca	0.5
Tex	0.8

Avg

Ca: 0.45
Ne: 0.15
Tex: 0.85

State	Score
Ca → 0.45	0.4
Ne → 0.15	0.1
Tex → 0.85	0.9
Ne → 0.15	0.2
Ca → 0.45	0.5
Tex → 0.85	0.8

缺点: 1) 依赖于y值

2) 对于 category 值，全都来自训练集，因此更容易出现 overfitting

4) 留一编码 (Leave-One-Out encoding)

改进了目标编码:

使用除了对应行 target 值外的其他同类 target 来计算。

State	Score
Ca	0.4
Ne	0.1
Tex	0.9
Ne	0.2
Ca	0.5
Tex	0.8

→ Avg. Exc': 0.5 →
→ Avg. Exc': 0.2 →
→ Avg. Exc': 0.8 →
→ Avg. Exc': 0.1 →
→ Avg. Exc': 0.4 →
→ Avg. Exc': 0.9 →

State	Score
0.5	0.4
0.2	0.1
0.8	0.9
0.1	0.2
0.4	0.5
0.9	0.8