五、特征工程

一、数值型数据处理

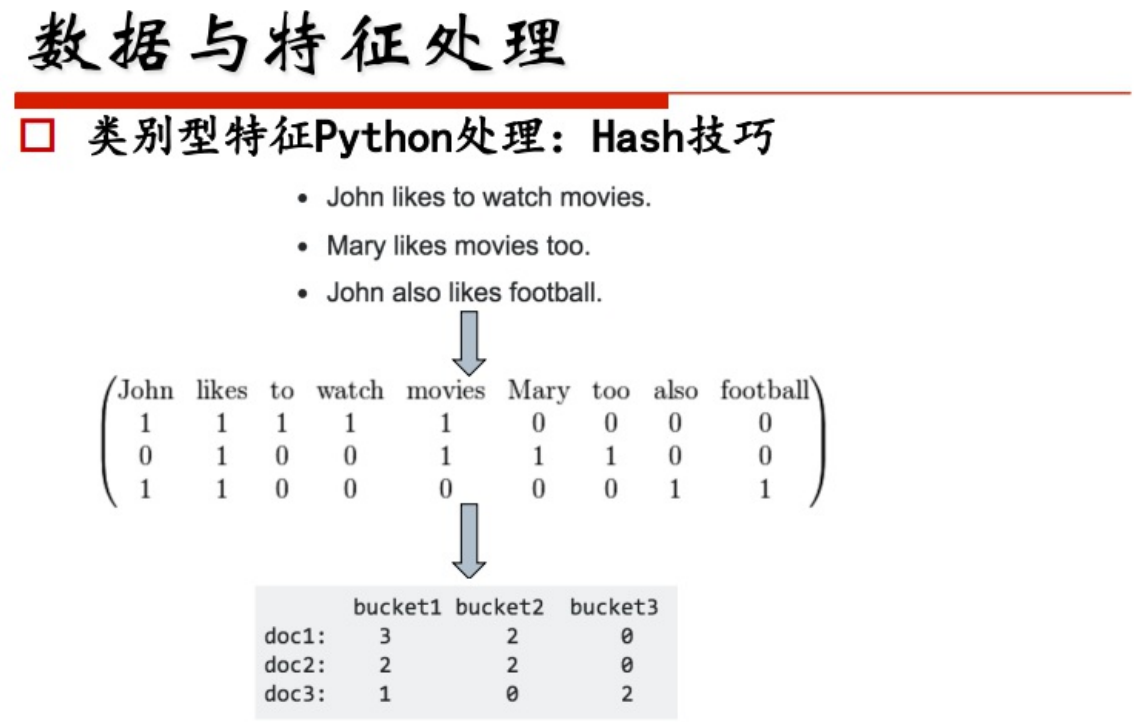
幅度调整（调整到一个数据范围内），归一化。离散化（连续值进行离散化）

二、类别型数据（类似于颜色，红色、蓝色之类）

1、one-hot编码

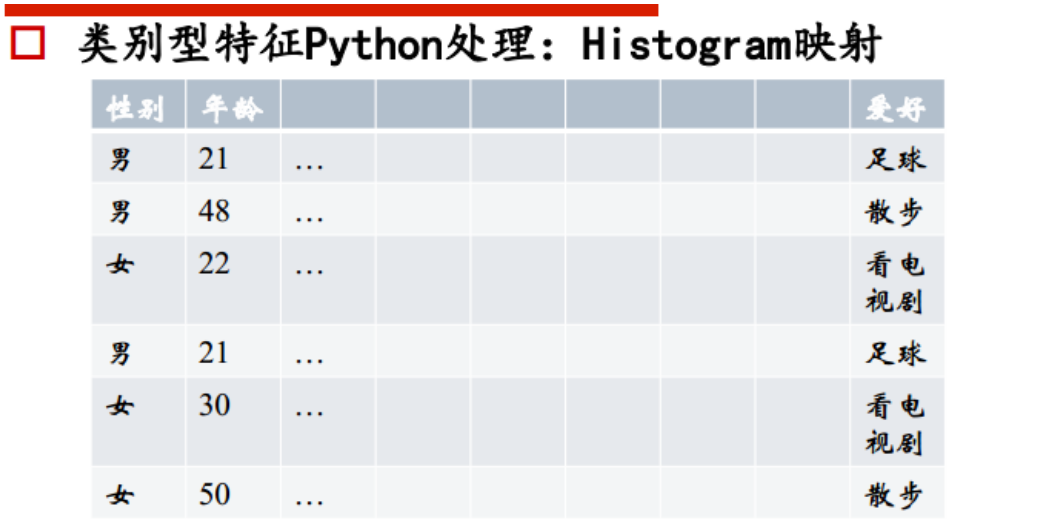
比如蓝色（1,0,0），红色（0,1,0），蓝色（0,0,1）

2、Hash与聚类方法



bucket表示的某个类别的词袋（比如体育，电影，李易峰）

3、Histogram映射



男：[1/3, 2/3, 0]； 女：[0, 1/3, 2/3]； 21：[1, 0, 0]；22：[0, 0, 1]…

[1/3,2/3,0]---[喜欢散步，喜欢散步，喜欢看电视剧]，把每一列特征拿出来，做对应target做统计，

三、时间型

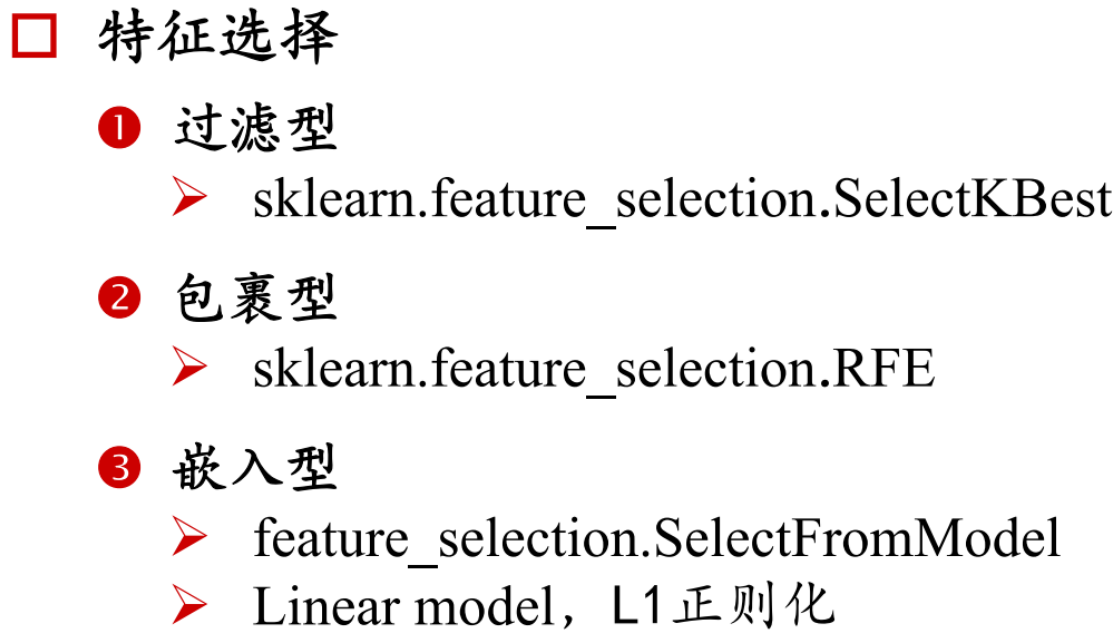
一周中的某一天，一年的那一周，一天中的第几个小时

四、文本型

词袋模型

td-idf

五、特征选择



过滤型是判断特征与标签的关联度

包裹型，简单的是计算LR，然后把前面参数θ较小的特征删除（不断的尝试删除）

嵌入型，利用L1正则项