

## 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

## 数据与特征处理

### □ 类别型特征Python处理：Hash技巧

- John likes to watch movies.
- Mary likes movies too.
- John also likes football.

A diagram showing the transformation of three sentences into a matrix. The sentences are: "John likes to watch movies.", "Mary likes movies too.", and "John also likes football.". Below them, a matrix is shown with columns corresponding to the words: John, likes, to, watch, movies, Mary, too, also, football. The matrix contains 1s where the word appears in the sentence and 0s otherwise.

	John	likes	to	watch	movies	Mary	too	also	football
1	1	1	1	1	1	0	0	0	0
2	0	1	0	0	1	1	1	0	0
3	1	1	0	0	0	0	0	1	1

A diagram showing the bucket counts for each document. The buckets are labeled bucket1, bucket2, and bucket3. The counts are: doc1: 3, 2, 0; doc2: 2, 2, 0; doc3: 1, 0, 2.

	bucket1	bucket2	bucket3
doc1:	3	2	0
doc2:	2	2	0
doc3:	1	0	2

bucket表示的某个类别的词袋（比如体育，电影，李易峰）

#### 3、Histogram映射

### □ 类别型特征Python处理：Histogram映射

性别	年龄							爱好
男	21	...						足球
男	48	...						散步
女	22	...						看电视剧
男	21	...						足球
女	30	...						看电视剧
女	50	...						散步

□ 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

## □ 特征选择

### ① 过滤型

➤ `sklearn.feature_selection.SelectKBest`

### ② 包裹型

➤ `sklearn.feature_selection.RFE`

### ③ 嵌入型

➤ `feature_selection.SelectFromModel`

➤ Linear model, L1正则化

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  $\theta$  较小的特征删除（不断的尝试删除）

嵌入型，利用L1正则项