

# 机器学习特征

## 目录

◆ 特征概念

◆ 特征编码

◆ 特征选择

## 特征概念

### 什么是特征

- ◆ 事物可供识别的特殊的征象或标志



可以用很多属性来描述一个西瓜，比如**色泽**、**根蒂**、**敲声**、**纹理**、**触感**等，这些描述事物的属性称为“**特征**”

# 特征选择

- ◆ 对于一个学习任务而言，有些特征可能很有用，另一些可能没什么用

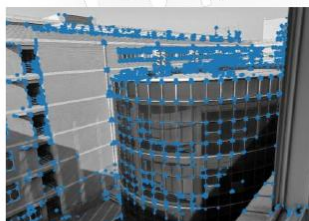


对当前任务有用的属性称为“相关特征”、没用的属性称为“无关特征”。从给定的特征集合中选出相关特征子集的过程称为“**特征选择**”。

有经验的人往往只需要看看**根蒂**、**听听敲声**就知道是否好瓜。

## 典型的图像特征

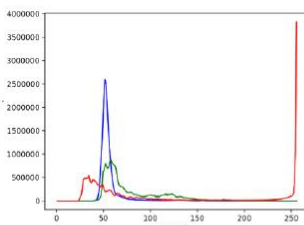
- ◆ 常用的特征有：Harris角点特征，Canny边缘特征，直方图特征等。



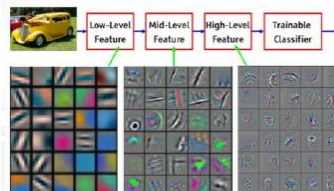
Harris角点



Canny算子



颜色直方图



深度特征

# 典型的文本特征

- ◆ 常用的特征有：词属性，词频TF-IDF，词向量，Bag of Words等。

king [ 0.30 0.70 ]  
man [ 0.20 0.20 ]  
woman [ 0.60 0.30 ]

词向量

	包含该词的文 档数 ( 亿 )	IDF	TF-IDF
中国	62.3	0.603	0.0121
蜜蜂	0.484	2.713	0.0543
养殖	0.973	2.410	0.0482

词频与逆文本频率指数

标记	含义	例子
ADJ	形容词	new, good, high, special, big, local
ADV	动词	really, already, still, early, now
CNJ	连词	and, or, but, if, while, although
DET	限定词	the, a, some, most, every, no
EX	存在量词	there, there's
FW	外来词	dolce, ersatz, esprit, quo, maitre
MOD	情态动词	will, can, would, may, must, should

词性



bags of words

## 机器学习数据库UCI

- ◆ UCI数据库是加州大学欧文分校(University of CaliforniaIrvine)提出的数据库，目前共有585个数据集，其数目还在不断增加。



Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 585 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#).

Supported By:  In Collaboration With: 

<http://archive.ics.uci.edu/ml/index.php>

# Iris数据库

- ◆ 可能是模式识别文献中最著名的数据库，数据集包含3个类，每个类有50个实例，每个类指的是一种鸢尾植物

## Iris Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Famous database, from Fisher, 1936



4个特征: 萼片长度(cm), 萼片宽度(cm), 花瓣长度(cm), 花瓣宽度(cm)

Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	3966673

# Adult数据库

- ◆ 从人口普查数据库中抽取，使用以下条件进行提取：(AGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)。

## Adult Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Predict whether income exceeds \$50K/yr based on census data. Also known as "Census Income" dataset.



Data Set Characteristics:	Multivariate	Number of Instances:	48842	Area:	Social
Attribute Characteristics:	Categorical, Integer	Number of Attributes:	14	Date Donated	1996-05-01
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits:	2140737

任务是预测确定一个人的年收入是否超过5万美元。

所采用的特征包括: 年龄、工作类型、教育程度、受教育时间、婚姻状况、职业、种族、性别、每周工作小时数、原籍、收入等。



## 特征编码

### 什么是特征编码

- ◆任务拿到的初始数据通常比较脏乱，可能会带有各种非数字特殊符号，需要将其转换为可计算的数字，采用编码量化等方法

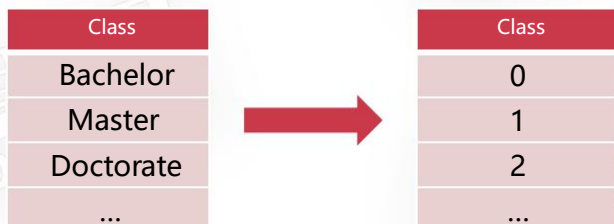
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K  
38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K  
53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K  
28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K  
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K  
49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, <=50K  
52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 45, United-States, >50K  
31, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White, Female, 14084, 0, 50, United-States, >50K  
42, Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 5178, 0, 40, United-States, >50K  
37, Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Black, Male, 0, 0, 80, United-States, >50K  
30, State-gov, 141297, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Islander, Male, 0, 0, 40, India, >50K  
23, Private, 122272, Bachelors, 13, Never-married, Adm-clerical, Own-child, White, Female, 0, 0, 30, United-States, <=50K  
32, Private, 205019, Assoc-acdm, 12, Never-married, Sales, Not-in-family, Black, Male, 0, 0, 50, United-States, <=50K  
40, Private, 121772, Assoc-voc, 11, Married-civ-spouse, Craft-repair, Husband, Asian-Pac-Islander, Male, 0, 0, 40, ?, >50K  
34, Private, 245487, 7th-8th, 4, Married-civ-spouse, Transport-moving, Husband, Amer-Indian-Eskimo, Male, 0, 0, 45, Mexico, <=50K

人口普查收入数据集

常见的编码方式有：序号编码、独热编码、标签编码、频数编码等

## 序号编码

- ◆ 序号编码 (Ordinal Encoding) , 通常用于处理类别间具有内在大小顺序关系的数据, 对于一个具有 $m$ 个类别的特征, 可以将其对应地映射到  $[0, m-1]$  的整数。



Class	Class
Bachelor	0
Master	1
Doctorate	2
...	...

例如收入数据集中的教育程度, 可以将“学士”、“硕士”编码成“0”和“1”, 因为它们内在就含有这样的逻辑顺序。

## 独热编码

- ◆ 独热编码 (one-hot encoding) , 又称一位有效编码, 使用 $N$ 位状态向量对 $N$ 个状态进行编码, 每个状态都有独立的位置, 并且只有一位有效。

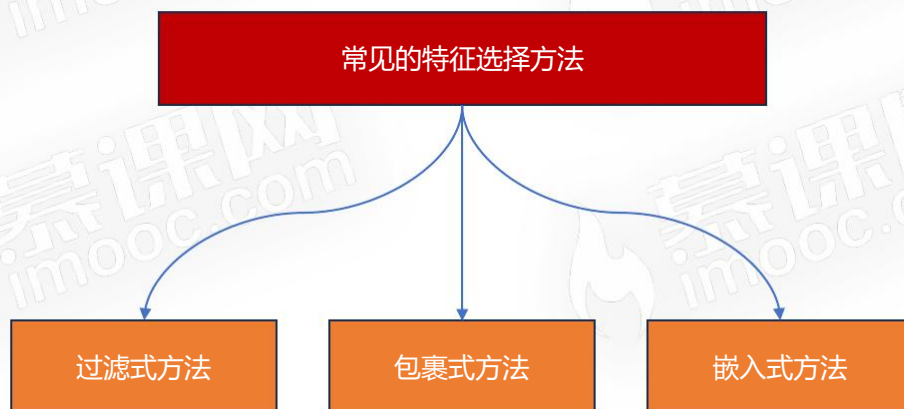
Class	One-hot
White	00001
Asian-Pac-Islander	00010
Amer-Indian-Eskimo	00100
Black	01000
Other	10000

通常用于处理类别间不具有大小关系的特征, 如收入数据集的种族。

# 特征选择

## 什么是特征选择

- ◆原始未经过滤过的特征可能包含很多无关特征，不同特征对问题的贡献大小程度也不相同，往往需要选取一个包含重要信息的特征子集。





## 过滤式特征选择方法

◆先对数据集进行特征选择，然后再训练学习器，特征选择过程与后续训练无关。

**方差选择法：**计算各个特征的方差，然后根据阈值，选择方差大于阈值的特征

**相关系数法：**计算各个特征对目标值的相关系数

**卡方检验法：**检验自变量对因变量的相关性。假设自变量有 $N$ 种取值，因变量有 $M$ 种取值，考虑自变量等于 $i$ 且因变量等于 $j$ 的样本频数的观察值与期望的差距

## 包裹式特征选择方法

◆直接以最终学习器的性能作为特征子集的评价准则，目的是为给定学习器选择最有利于其性能的特征子集。

**递归特征消除法：**使用一个基模型进行多轮训练，每轮训练后，根据权值系数消除若干特征，再基于新的特征集进行下一轮训练。

## 嵌入式特征选择方法

- ◆将特征选择过程与学习器训练过程融为一体，两者在同一个优化过程中完成

**基于惩罚项的特征选择法：**使用L1等惩罚项对不同维度的特征进行惩罚，除了筛选出特征外，也进行了降维。

**基于树模型的特征选择法：**使用决策树、随机森林、Boosting、XGBoost等算法进行特征选择。

**下次预告：机器学习模型种类**