机器学习的数据：文件csv

mysql:1、性能瓶颈，读取速度

2、格式不太符合机器学习要求数据的格式

pandas:读取工具

真正的多线程

4个线程

1、历史遗留问题

numpy:释放了GIL

cpython

jpython

特征值+ 目标值

特征

身高 体重 皮肤颜色 头发长度 目标值

列索引 男

1 女

2

3

文本

dataFrame 缺失值，数据转换

机器学习： 重复值? 需要进行去重 特征抽取：文本，字符串

特征工程

numpy　　scipy　　sparse矩阵

```
['city=上海', 'city=北京', 'city=深圳', 'temperature']
```

节约内存，方边读取处理

```
(0, 1)    1.0
(0, 3)    100.0
(1, 0)    1.0
(1, 3)    60.0
(2, 2)    1.0
(2, 3)    30.0
```

字典数据抽取：把字典中一些类别数据，分别进行转换成特征

ndarray　　二维数组

```
/Users/huxinghui/virtualenv/mlt3/b
[[   0.    1.    0.  100.]
 [   1.    0.    0.   60.]
 [   0.    0.    1.   30.]]

Process finished with exit code 0
```

数组形式，有类别的这些特征
先要转换字典数据

One-hot编码

对于单个英文字母不同：没有分类依据

文本特征抽取： Count

tf * idf

词语占比

共享

证券

文章类型？？？

Tf:term frequency:词的频率

idf:逆文档频率inverse document frequency

文本分类

情感分析

所以， 我们，明天

朴素贝叶斯

log(数值): 输入的数值越小，结果越小

重要性程度

出现的次数

log(总文档数量/该词出现的文档数量)

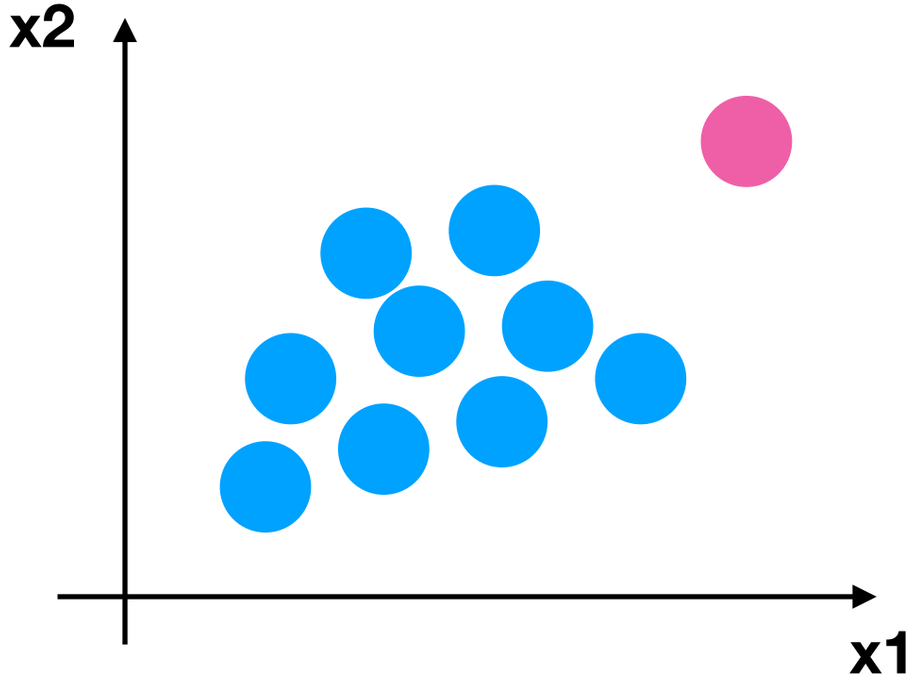
公式： $X' = \frac{x - min}{max - min}$ $X'' = X' * (mx - mi) + mi$

注：作用于每一列，max为一列的最大值，min为一列的最小值,那么X"为最终结果，mx，mi分别为指定区间值默认mx为1,mi为0

里程数	公升数	消耗时间比	评价
14488	7.153469	1.673904	smallDoses
26052	1.441871	0.805124	didntLike
75136	13.147394	0.428964	didntLike
38344	1.669788	0.134296	didntLike
72993	10.141740	1.032955	didntLike
35948	6.830792	1.213192	largeDoses
42666	13.276369	0.543880	largeDoses
67497	8.631577	0.749278	didntLike
35483	12.273169	1.508053	largeDoses
50242	3.723498	0.831917	didntLike

(72993-35948)^2 + (10.14-6.8)^2+(1.0-1.21)^2

目的：是的某一个特征对最终结果不会造成更大影响



x1,x2

异常点对最大值最小值影响太大

2、公式：$X' = \frac{x - \text{mean}}{\sigma}$

注：作用于每一列，mean为平均值，σ为标准差

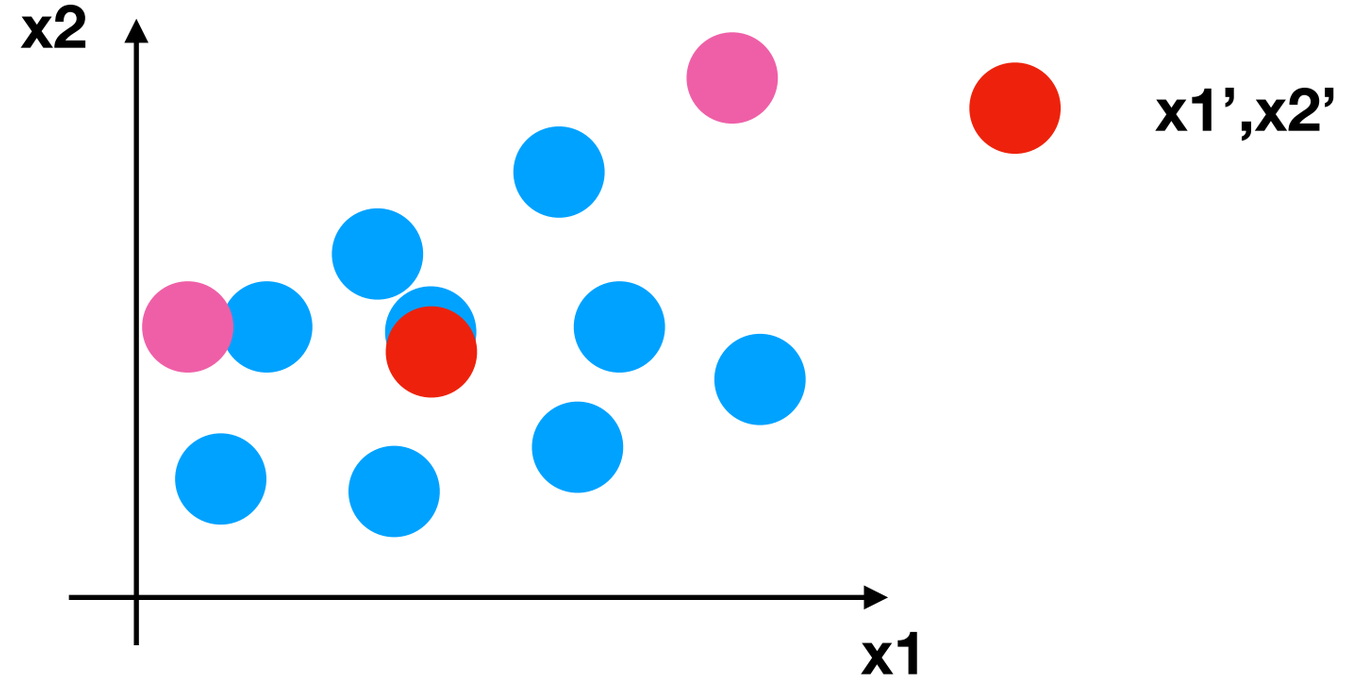
var成为方差，$var = \frac{(x1-mean)^2+(x2-mean)^2+\cdots}{n(\text{每个特征的样本数})}$，$\sigma = \sqrt{var}$

variance

其中： 方差(考量数据的稳定性)

异常点

特征1	特征2	特征3	特征4
90	2	10	40
60	4	15	45
75	3	13	46



方差：0

所有这个特征所有值都一样

pandas:dropna

fillna 数据当中的缺失值：np.nan

replace(“?”, np.nan)

维度（数组的维度）

降维：维度：特征的数量

特征1	特征2
2	20
2	20
	19
2	20
2	19
2	20
2	20
	20

贷款额度

方差大小：考虑所有样本这个特征的数据情况

PCA：　特征数量达到上百的时候

考虑数据的简化

数据也会改变，特征数量也会减少

特征1	特征2	特征3	特征50，。。。。特征100
1			2
2			4
3			6
4			8
5			9
6			12
7			13
8			15

*n_components:*小数　　**0~1　90%　　90~95%**

整数　**减少到的特征数量**

instacart:把用户分成几个类别　　用户—购买的物品类别

购买的物品类别

用户1
用户2
用户3

数据：

- products.csv　　商品信息
- order_products__prior.csv 订单与商品信息
- orders.csv　　用户的订单信息
- aisles.csv　　商品所属具体物品类别

1、合并各张表到一张表当中

pd.merge()

2、建立一个类似行，列数据

prior: product_id,order_id
products:product_id, aisle_id
orders:order_id,user_id
aisles:aisle_id,aisle

交叉表（特殊的分组表）

机器学习算法分类

监督学习：特征值+目标值

非监督学习：特征值 1000个样本

- 监督学习（预测）
  - 分类　k-近邻算法、贝叶斯分类、决策树与随机森林、逻辑回归、神经网络
  - 回归　线性回归、岭回归
  - 标注　隐马尔可夫模型　(不做要求)
- 无监督学习
  - 聚类　k-means

分类：　目标值离散型

回归：目标值连续型

1、公司本身就有数据

数据

2、合作过来数据

3、购买的数据

建立模型：根据数据类型划分应用种类

1、原始数据明确问题做什么

2、数据的基本处理：pd去处理数据（缺失值，合并表。。。。。）

3、特征工程 （特征进行处理） (重要)

分类：

回归：

模型：算法 + 数据

2、特征工程

4、找到合适算法去进行预测

1、换算法 参数

没有合格

5、模型的评估，判定效果

上线使用 以API形式提供