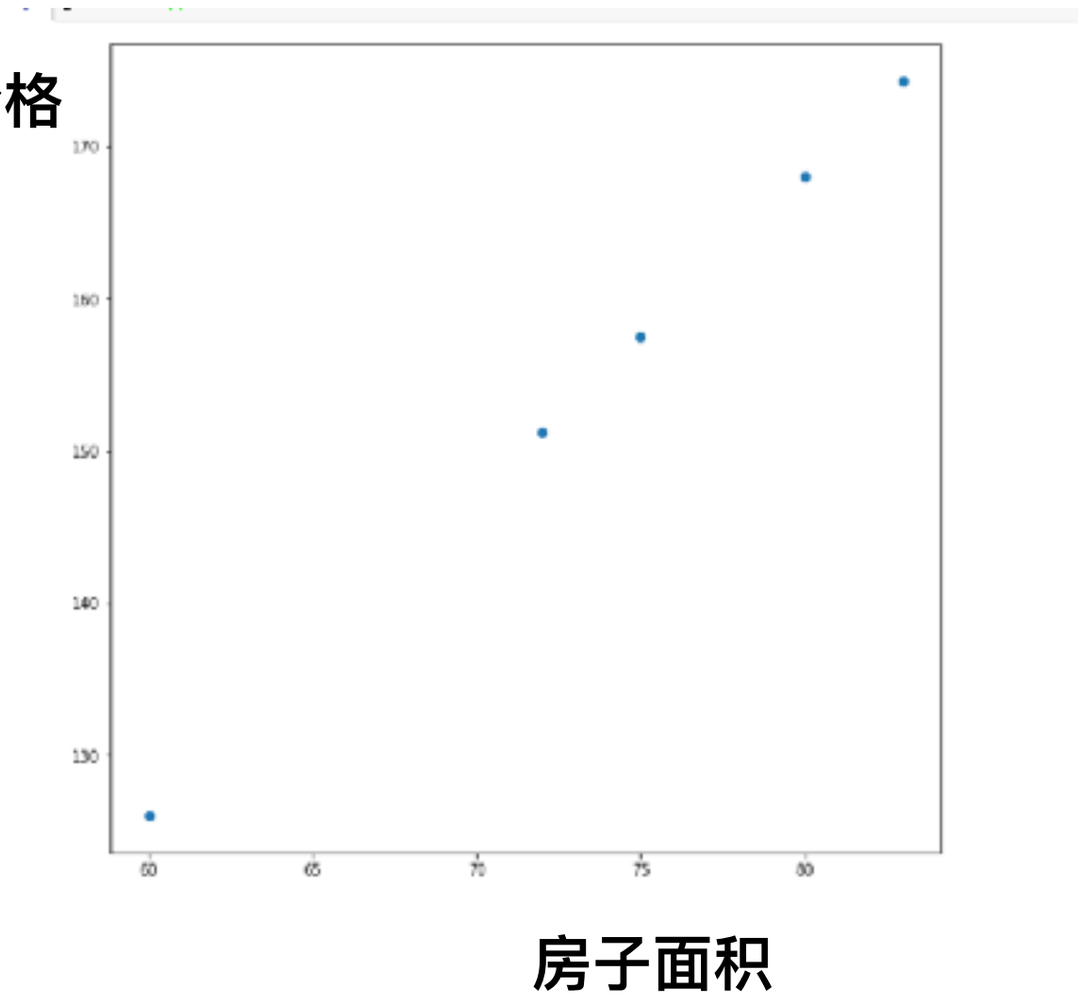
分类：离散型　　0 ，1，　2, 18

回归：目标值连续　　225.35　　100.234

回归算法

平时成绩	考试成绩	期末成绩
0.3	0.7	

线性回归：寻找一种能预测的趋势



线性关系：二维：直线关系

三维：特征，目标值 ， 平面当中

线性关系定义：

$$y = w x + b$$ b:偏置

加b：为了是对于单个特征的情况更加通用

多个特征：w1房子面积+w2房子位置 + b

通用公式：$h(w) = w_0 + w_1x_1 + w_2x_2 + \cdots = w^Tx$

其中w, x为矩阵：$w = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix}, x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$

属性和权重一种组合来预测结果

矩阵：大多数算法计算基础

数组 **矩阵**

0维 5

1维 [5,2,3,45,676]

2维 [[5,2,3,45,676]] **必须是二维**

3维
```
[
[[5,2,3,45,676]],
[[5,2,3,45,676]]
]
```

矩阵乘法： 满足了特定运算需求

(m行，l列) * (l行，n列) = （m行， n列）

数组的运算：加法，乘法

numpy:ndarray

特征值	权重	目标值
[[1,2,3,4] [5,6,7,8]]	[[2],[2],[2],[2]]	一个样本应该是一个值
(100,4)	(4,1)	(100,1)

回归：迭代的算法

神经网络

预测时候有差距

《统计学习方法》

算法

线性回归

逻辑回归

策略(损失函数)

误差平方和

最小二乘法

对数似然损失

优化

正规方程

梯度下降学习率

梯度下降

回归：知道误差，也去不算减少

损失函数最小　寻找最优化的W值

算法的自我学习的过程

求解：$w = (X^T X)^{-1} X^T y$

X为特征值矩阵，y为目标值矩阵

[[1,2,3,4]
[0.2,1.0,5,6]
[0.5,0.49,3,4]
[6,5,8,9]]

w:4个

[[2]
[1.5]
[2.1]
[3.4]]

scikit-learn:优点：封装好，建立模型简单，预测简单

缺点：算法的过程，有些参数都在算法API内部优化

tensorflow:封装高低，　自己实现线性回归，学习率等等

0.18　　二维，1维都可以

0.19　转换器，estimator　要求数据必须是二维

线性回归：线性关系数据　系数

非线性关系　系数

10个特征　　预测值　模型复杂的原因：数据的特征和目标值之间的关系

不仅仅是线性关系

w

根据结果现象判断：欠拟合，过拟合

交叉验证：训练集结果：表现不行　　结果99%，2.0

测试集：表现不行　　结果89%，10.0

欠拟合　　过拟合

特征选择：过滤式：低方差特征

嵌入式：正则化，决策树，神经网络

回归：解决过拟合的方式

线性回归：LInearRegression　容易出现过拟合，为了把训练集数据表现更好

L2正则化：Ridge:岭回归　带有正则化的线性回归　解决过拟合

逻辑回归：线性回归的式子作为的输入　　二分类　　也能得出概率值

点击　　0.001　　0.01

广告点击

没点击

线性回归的输入 ——sigmoid——> 分类　[0, 1]　概率值

完整的损失函数：

$$cost(h_\theta(x), y) = \sum_{i=1}^{m} -y_i log(h_\theta(x)) - (1 - y_i) log(1 - h_\theta(x))$$

阈值：0.5

[样本1， 样本2，样本3，样本4] **逻辑回归预测 [0.6 0.1 0.51 0.7]]** **[4 2 2 4]**

属于4的概率值

更新权重

四个损失值相加

1log(0.6) + 0log(0.1)+ 0log(0.51)+ 1log(0.7) **信息熵**

9.班级分布：

良性：458（65.5%）
恶性：241（34.5%）

哪一个类别少，判定概率值是值得这个类别 恶性 正例

良性 反例

损失函数：均方误差　（不存在多个局部最低点）　只有一个最小值

对数似然损失：　多个局部最小值　　目前解决不了的问题

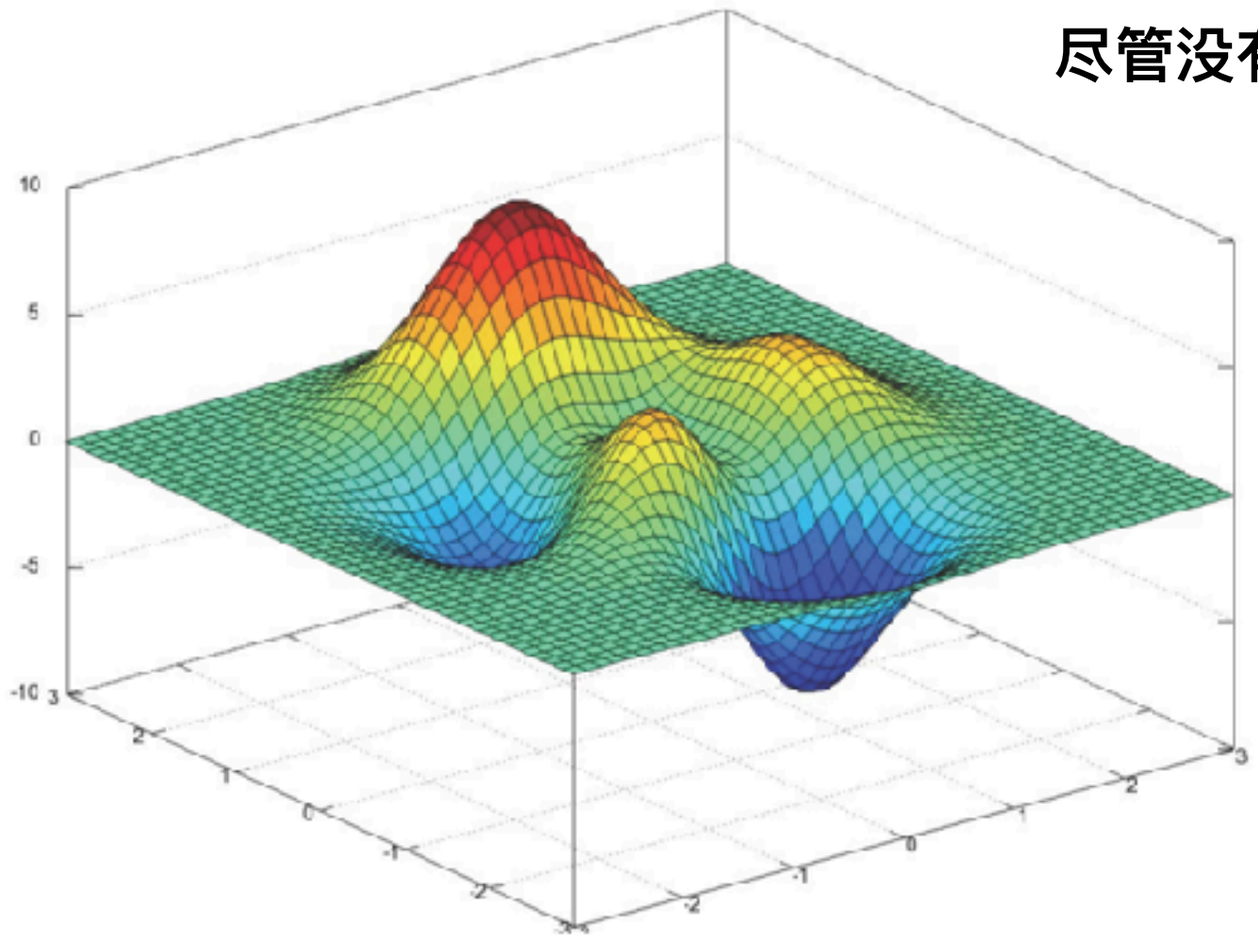
梯度下降求解

1、多次随机初始化，多次比较最小值结果

2、求解过程当中，调整学习率

尽量改善

尽管没有全局最低点，但是效果都是不错的



判别模型

生成模型

先验概率 P(c)

逻辑回归

朴素贝叶斯

解决问题

二分类

多分类问题

P(f1,f2..|c)P(c)

应用场景

癌症，二分类需要概率

文本分类

参数

正则化力度

没有

得出的结果都有概率解释

隐马尔可夫模型

k-近邻，决策树，随机森林，神经网络

1000个数据

x1,x2

聚类

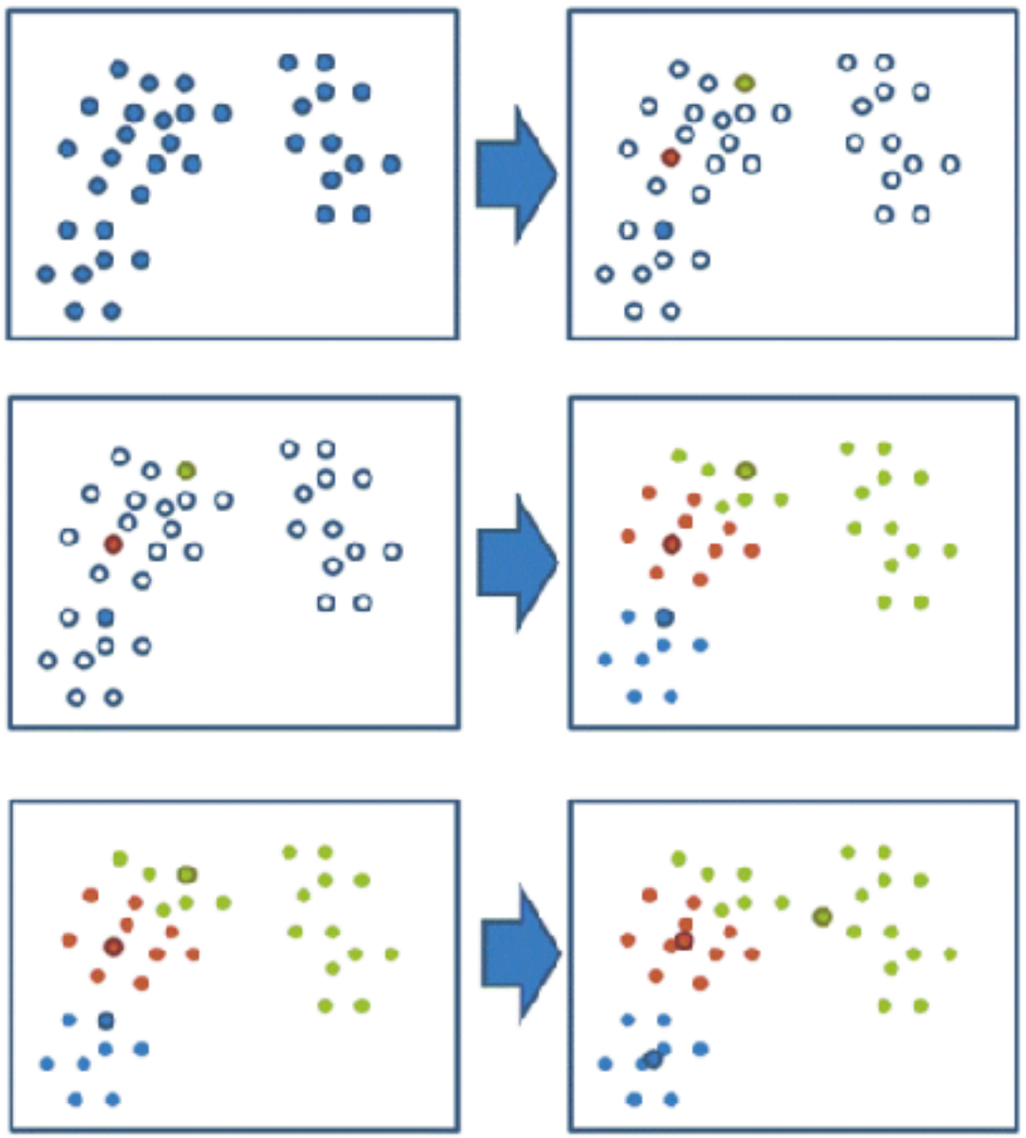
K:把数据划分成多少个类别

知道类别的个数

不知道类别个数

超参数

k=3



1、随机在数据当中抽取三个样本，当做三个类别的中心点(k1,k2,k3)

2、计算其余的点分别到这三个中心点的距离，每一个样本有三个距离(a,b,c)，从中选出距离最近的一个点作为自己的标记

形成三个族群

3、分别计算这三个族群的平均值，把三个平均值与之前的三个旧中心点进行比较

如果相同：结束聚类

如果不相同：把这三个平均值当做新的中心点，重复第二步

绿1 (x1,x2)

绿2(x1',x2')

平均值（x1平，x2平）

聚类　做在分类之前

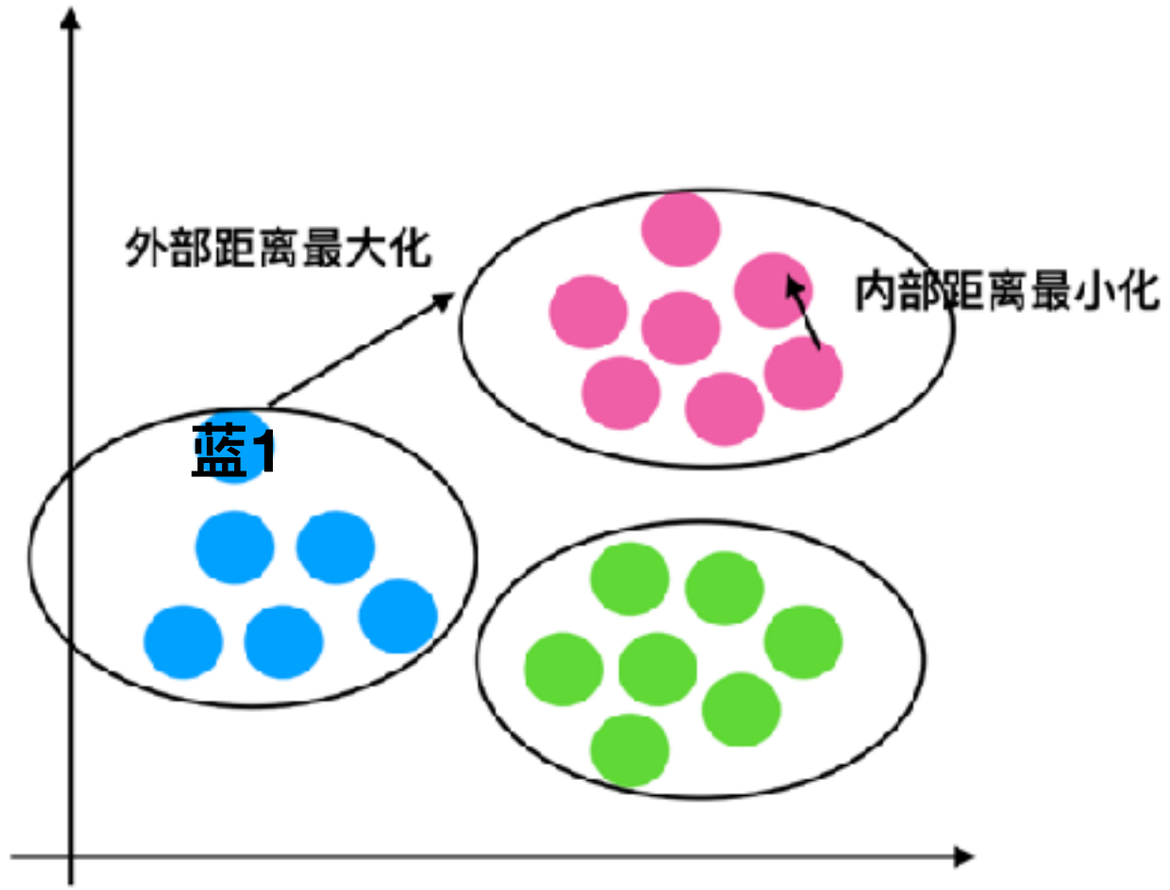
聚类评估标准

轮廓系数：

计算公式：$sc_i = \frac{b_i - a_i}{\max(b_i, a_i)}$

注：对于每个点i为已聚类数据中的样本，b_i为i到其它族群的所有样本的距离最小值，a_i为i到本身簇的距离平均值

最终计算出所有的样本点的轮廓系数平均值



轮廓系数　　对于每一个样本

1、计算蓝1到自身类别的点距离的平均值a_i

2、计算蓝1分别到红色类别，绿色类别所有的点的距离，求出平均值
b1, b2，取其中最小的值当做b_i

极端：

b_i>>a_i: 1　　完美

a_i>> b_i: -1　　最差

蓝1：轮廓系数　[-1, 1]　　0.1