1.knn算法（k近邻算法）

基于欧式距离公式



工作原理：

1. 假设存在一个带有标签的样本数据集（训练样本集），其中包含了每条数据与所属分类的对应关系。
2. 输入没有标签的待测数据后，将待测数据的每个特征与样本集中的数据对应特征进行比对。
3. 计算待测数据与训练数据集中每条数据的距离。
4. 对求得的所有距离进行排序（从小到大，越小表示越相似）。
5. 取前k个样本数据对应的分类标签。
6. 求k个数据中出现次数最多的标签作为待测数据的标签。

算法优缺点：

优点: 精度高、对异常值不敏感、无数据输入假定

缺点: 计算复杂度高、空间复杂度高

适用数据范围: 数值型和标称型

项目1：优化约会网站的配对效果

输入数据为：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 每年获得的飞行常客里程数 | 玩视频游戏所耗时间百分比 | 每周消费的冰淇淋公升数 | 样本分类 |
| 1 | 400 | 0.8 | 0.5 | 1 |
| 2 | 134 000 | 12 | 0.9 | 3 |
| 。。。。 | 。。。。 | 。。。。 | 。。。 | 。。 |

共1000多条数据；样本分类1代表工作日与魅力一般的人约会，2代表，3代表不喜欢的人则直接排除掉。

选择百分之十的数据作为测试数据，百分之九十的数据作为训练数据。

100条数据里错误率为百分之五。

项目2：手写数字识别系统

输入数据为：需要识别的数字是存储在文本文件中的具有相同的色彩和大小: 宽高是 32 像素 \* 32 像素的黑白图像。

先将已有的32\*32的01矩阵训练数据转化为1\*1024的01矩阵。

讲测试数据与训练数据进行对比，取个k个距离最近的数据中数额最高的那一个标签。

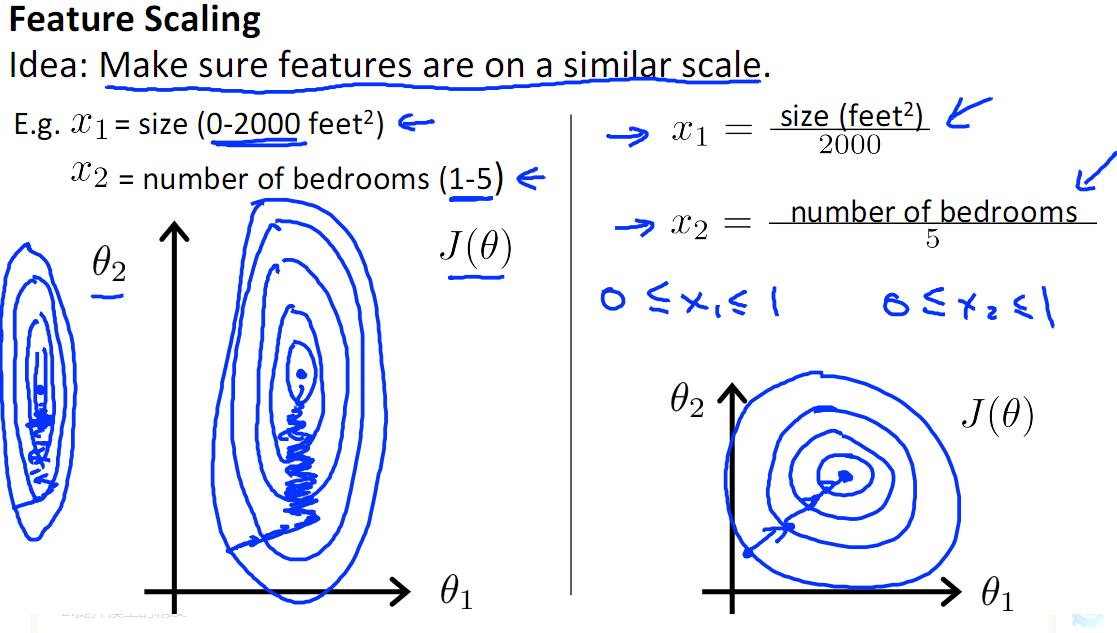
存在疑惑：如何构建一个32像素\*32像素的黑白数字图像，并转换为32\*32的01矩阵？

2.归一化数据

不同评价指标往往具有不同的量纲和量纲单位，这样的情况会影响到数据分析的结果，为了消除指标之间的量纲影响，需要进行数据标准化处理（如果上表数据中存在有超过10000的数据）。归一化处理后的数据各指标都会处于同一量级上。



如下图，x1的取值为0-2000，而x2的取值为1-5，假如只有这两个特征，对其进行优化时，会得到一个窄长的椭圆形，导致在梯度下降时，梯度的方向为垂直等高线的方向而走之字形路线，这样会使迭代很慢，相比之下，右图的迭代就会很快



梯度下降的作用？

损失函数就是一个以算法为自变量，输出值为误差值的函数。而梯度下降就是为了找到误差值最小的时候算法取的参数。而什么是损失函数，损失函数也可以叫误差函数，机器学习中有一种算法是生成一条曲线来拟合现有数据，也可以叫分类现有数据，这条曲线可以对未来的数据进行预分类。但是预测毕竟是预测不可能做到百分百正确，而其中会存在误差。一个算法不同参数会有不同的曲线，而梯度下降的目的就是找到误差值最小的时候这个算法的参数。而归一化数据可以让梯度下降更快。这是其中一个好处。

第二个好处是提升预测模型的精度，比如算法为欧式距离计算时，上图的x2取值范围远超x1，导致x2对结果的影响远超x1，会照成预测精度的损失，这是不合理的。而归一下数据后，所有特征数据都会处以一个量级之下，大大减少了精度损失。