

图像的特征：

1. 特征工程：图像转化为像素值

2. rotation-invariant (旋转不变)

- 颜色特征 (Color Histogram)
- SIFT (Scale-invariant feature transform)
- HOG (Histogram of Oriented Gradient)

颜色特征：常用颜色直方图-R,G,B

SIFT (尺度不变特征转换)：局部特征，寻找图片中的拐点这类的关键点，处理得到一个SIFT向量。

HOG (方向梯度直方图)：在单元格上，通过计算和统计图像局部区域的梯度方向直方图来构建特征。

PCA(Principal Component Analysis)：一种无监督的学习方法，可以把高维的向量映射到低维的空间里。它的核心思路是对数据做线性的变换，然后在空间里选择信息量（特征值影响）最大的Top K维度作为新的特征值。

KNN进阶

一、缺失值的处理：

删除法：

1. 第一种删除法是把相应的属性全部删掉，也就是删掉整个列。对于某一特征缺失较多的时候适用。

2. 第二种删除法是删除相应的记录。只要一个记录里包含了缺失值，就丢弃掉此记录，删除行。对于维度很多的数据，假如大量的维度都存在缺失值，这个方法很容易丢弃掉大量的样本。

填补法：

用新的值来填补缺失值。

1. 平均值/中位数填补法：用特征的平均值来填补。

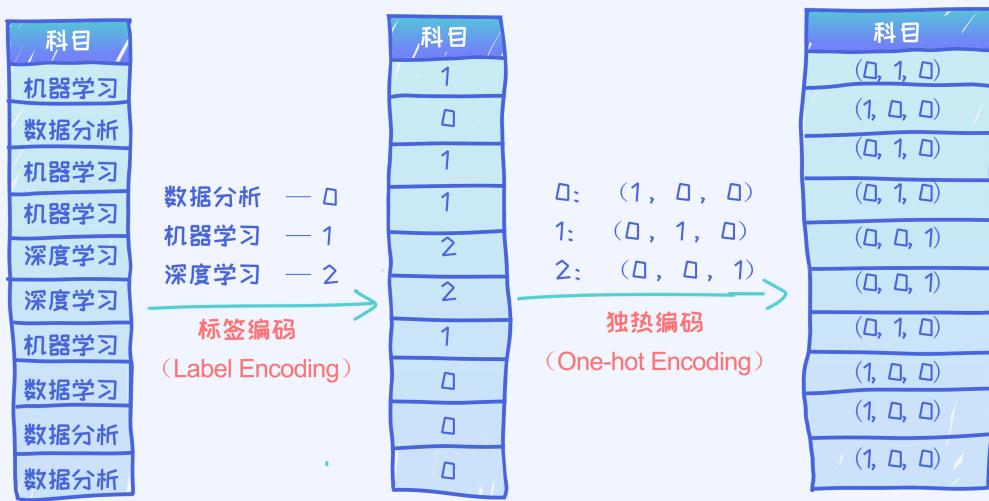
二、特征编码(feature encoding)

将字符串等类别特征转换成数值类型的过程。

独热编码(one-hot encoding): 在标签特征的基础上需要创建一个向量。这个向量的长度跟类别种类的个数等同的，另外，除了一个位置是1，其他位置均为0，1的位置对应的是相应类别出现的位置。

标签编码(label encoding)

特征编码 - 独热编码 (One-hot Encoding)



数值型变量处理：变量的离散化操作-连续性特征的离散化操作可以**增加模型的非线性型**，同时也可以有效地**处理数据分布的不均匀**的特点。

顺序变量：很少使用独热编码，因为这样就失去了大小关系。

三、KNN的复杂度分析以及KD树

提升kNN的搜索效率

1. 对于每一个类别，我们只选择具有代表性的几个样本，之后的kNN搜索就只依赖于这些样本
2. 使用近似kNN算法，但这会损失一些精度
3. 使用kD树，但一般只能用在低维度的空间

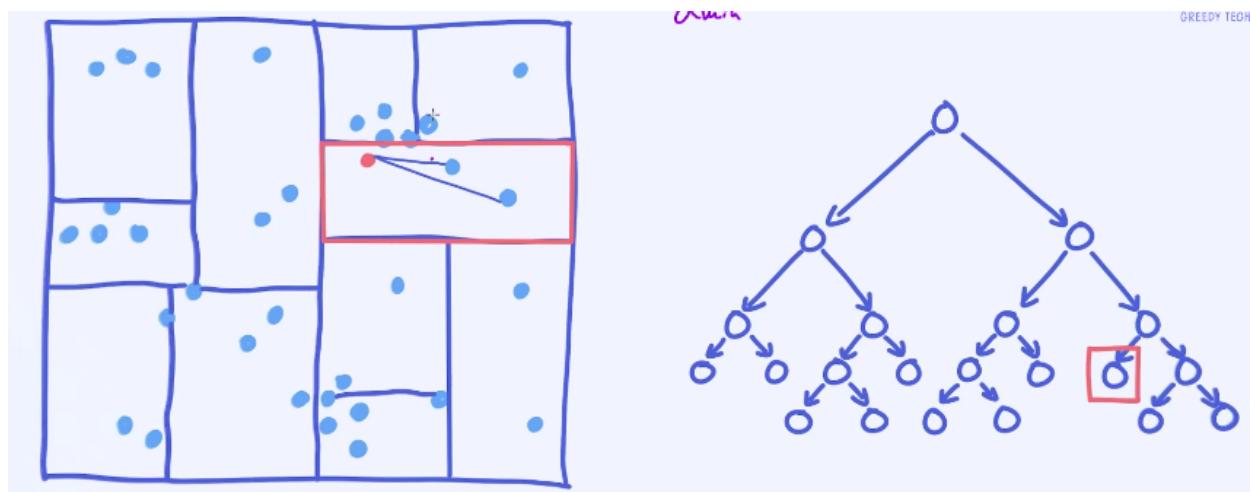
近似KNN：搜索过程中会做一些近似运算来提升效率，但同时也会牺牲一些准确率。

KD树 (k-dimensional tree) : kd树是二叉树，表示对k维空间的一个划分(partition)

- 距离最短的点不一定落在跟预测样本同一个区域。
- 为了保证能够找到全局最近的点，我们需要适当去检索其他区域里的点，这个过程也叫作**Backtracking (回溯)**。
- 随着特征维度的增加，KD树的搜索时间复杂度会指数级增加，因此不适合在高维空间里。主要是回溯会带来高频的搜索查找，造成维数灾难 (curse of dimension)。

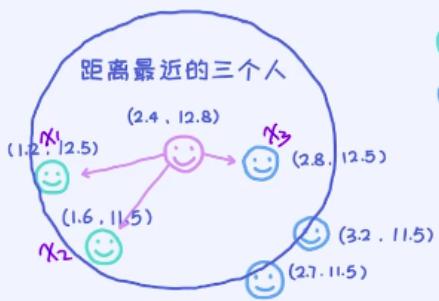
KD树总结以及改进：<https://www.jianshu.com/p/abcaaf754f92>

应用场景：地图搜索最近的店



带权重的KNN: 思路是跟预测目标越近的样本给与了更高的权重，离预测目标越远的样本就越低的权重。最后做了加权平均。

带权重的 KNN k=3



预测

$$Pr(y|x) = \frac{\sum_{i=1}^n w(x, x_i) \delta(y, y_i)}{\sum_{i=1}^n w(x, x_i)}$$

$$\delta(y, y_i) = \begin{cases} 1 & y = y_i \\ 0 & y \neq y_i \end{cases}$$

Offer

No Offer

权重计算

$$w(x, x_i) = \exp(-\lambda |x - x_i|_2^2)$$

$$w(x, x_1) = \exp(-\lambda (1.2^2 + 12.5^2))$$

$$w(x, x_2) = \exp(-\lambda (1.6^2 + 11.5^2))$$

$$w(x, x_3) = \exp(-\lambda (2.4^2 + 12.8^2))$$

$$Pr(y=\text{offer}|x) = \frac{w(x, x_1) + w(x, x_2)}{w(x, x_1) + w(x, x_2) + w(x, x_3)}$$

$$Pr(y=\text{No offer}|x) = \frac{w(x, x_3)}{w(x, x_1) + w(x, x_2) + w(x, x_3)}$$

条件概率

λ : 超参数, 人工定义值

距离: 计算目标参数与样本参数之间的距离为欧式距离

δ 函数: 目标与样本之间类别是否一致