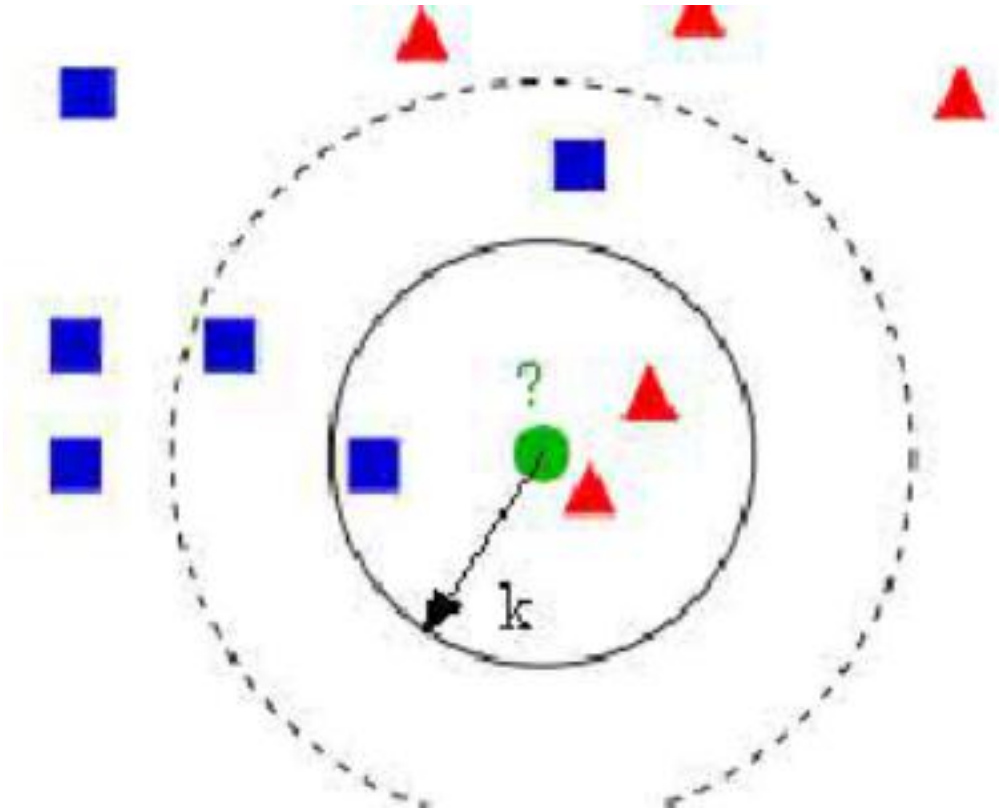
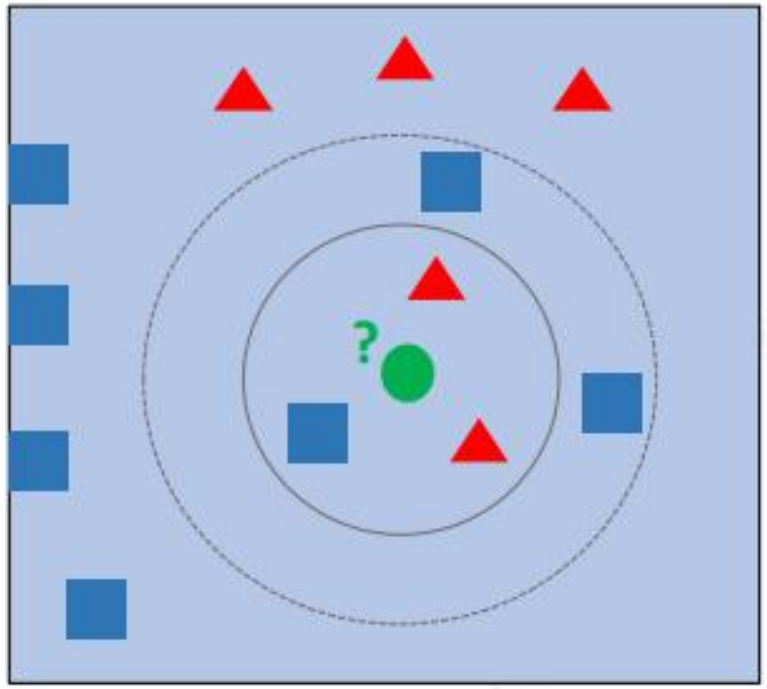
KNN算法



近邻分类思想

- KNN（k-Nearest Neighbor)分类算法是数据挖掘分类技术中较简单的方法之一。
- 所谓k最近邻，就是k个最近的邻居的意思，说的是每个样本都可以用它最接近的k个邻居来代表。



例如，上图中，绿色圆要被决定赋予哪个类，是红色三角形还是蓝色四方形？如果

K=3，由于红色三角形所占比例为2/3，绿色圆将被赋予红色三角形那个类，如果

K=5，由于蓝色四方形比例为3/5，因此绿色圆被赋予蓝色四方形类。

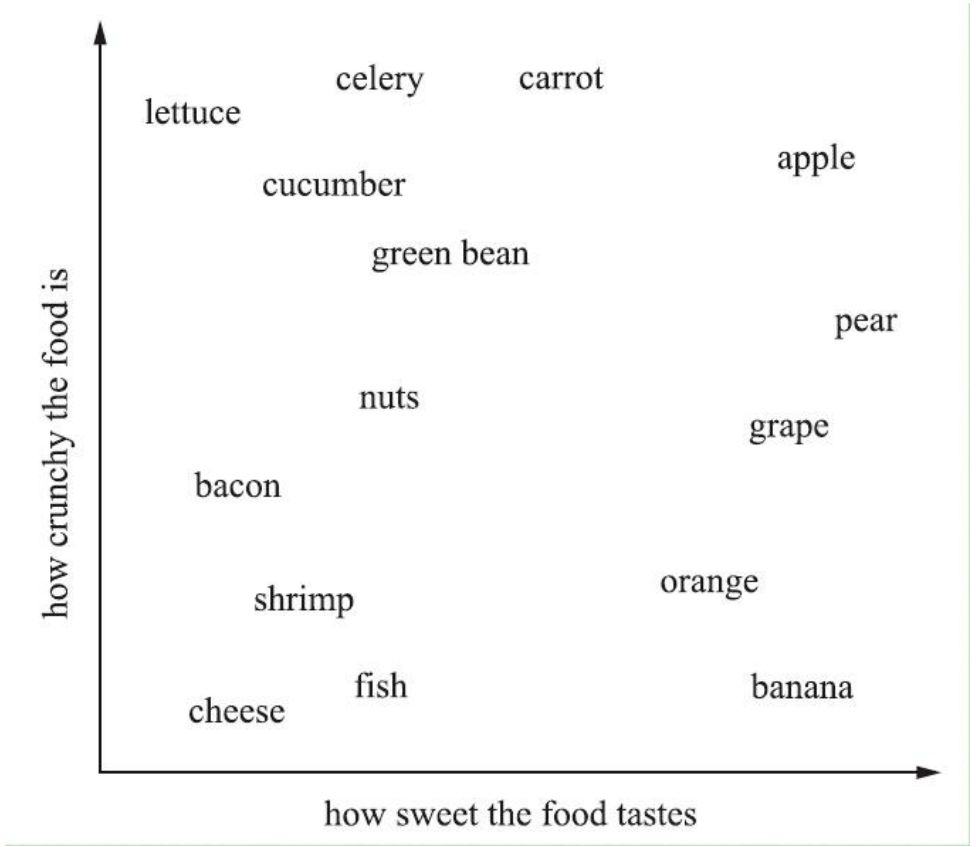
一个案例来了解kNN

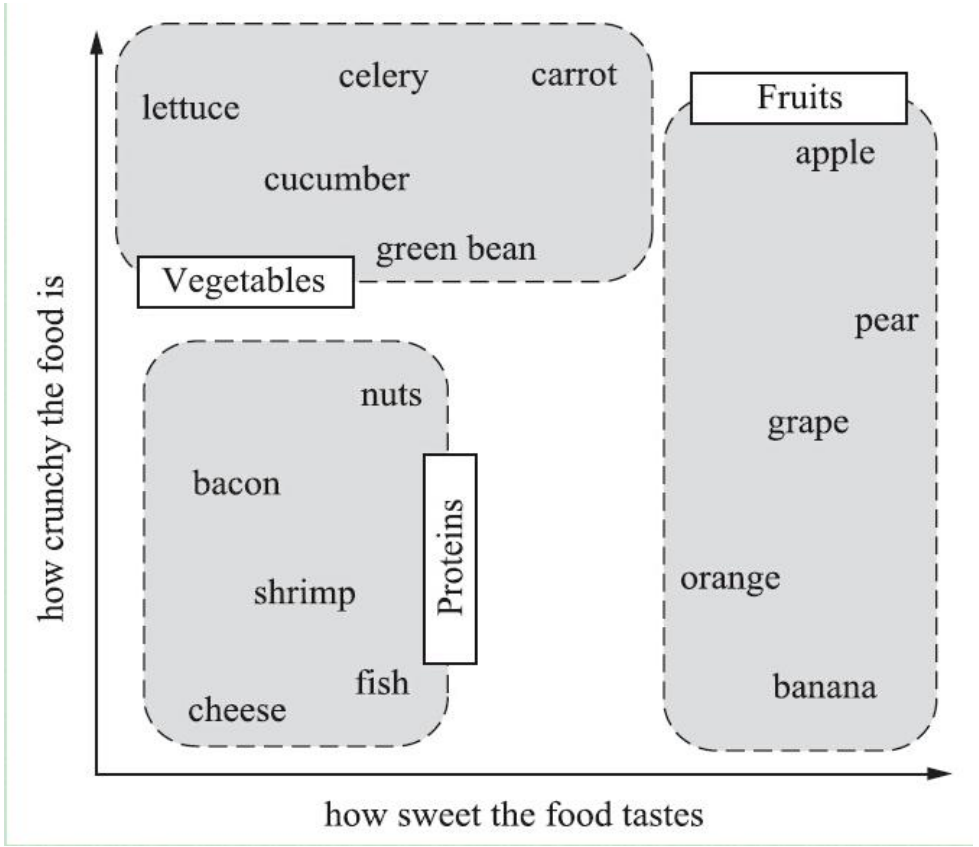
ingredient	sweetness	crunchiness	food
apple	10	9	fruit
bacon	1	4	protein
banana	10	1	fruit
carrot	7	10	vegetable
celery	3	10	vegetable
cheese	1	1	protein

- 上面对多种食物提供两个特征，一个特征是对配料有多脆的度量(crunchiness),从1~10；第二个特征是对配料有多甜的度量（sweetness）,从 1~10。
- 我们标记配料为3中类型之一：fruit（水果）、vegetable（蔬菜）或者protein（蛋白质）
- 上表即为数据表的前几行

一个案例来了解kNN

我们绘制二维数据的散点图，维度X表示配料的甜度（sweetness），维度y表示配料的脆度（crunchiness），散点图如下：





一个案例来了解kNN

- 西红柿是属于哪类呢

KNN参数说明

IT在线教育领导品牌
EDUCATION TO CREATE A BRIGHT FUTURE

参数	KNeighborsClassifier	KNeighborsRegressor
weights	样本权重，可选参数: uniform(等权重)、distance(权重和距离成反比，越近影响越强)；默认为uniform	
n_neighbors	邻近数目，默认为5	
algorithm	计算方式，默认为auto，可选参数: auto、ball_tree、kd_tree、brute；推荐选择kd_tree	
leaf_size	在使用KD_Tree的时候，叶子数量，默认为30	
metric	样本之间距离度量公式，默认为minkowski（闵可夫斯基）；当参数p为2的时候，其实就是欧几里得距离	
p	给定minkowski距离中的p值	