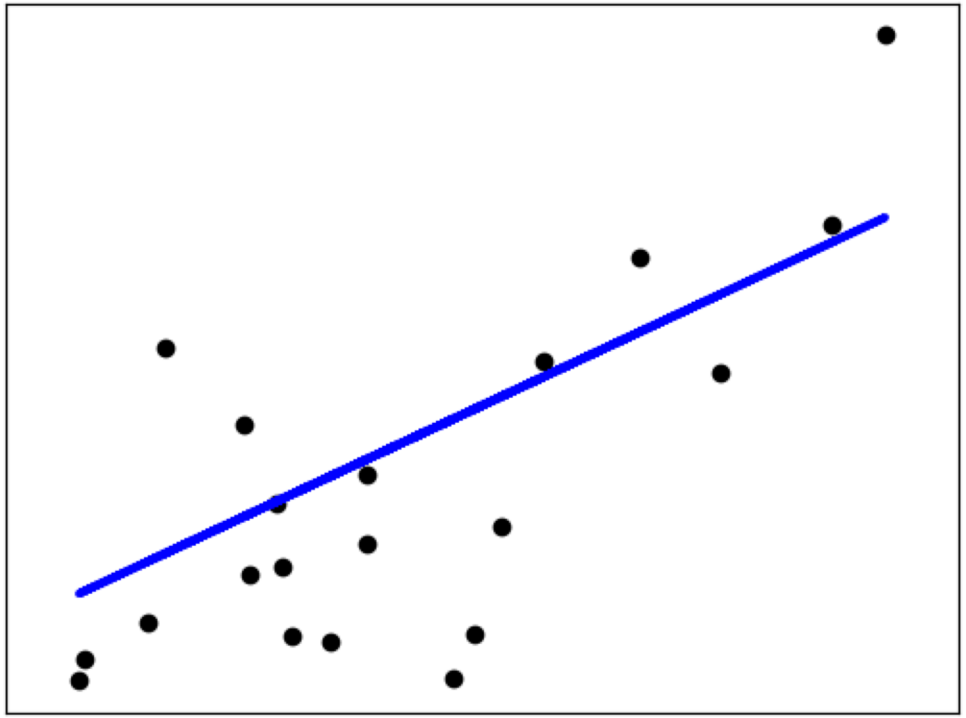
监督学习：回归问题

回归问题——线性回归

基本思路：使用线性函数拟合输入特征和预测结果

一元线性回归：
$f(x) = ax + b$
二元线性回归：
$f(x_1, x_2) = a_1x_1 + a_2x_2 + a_0 = \boldsymbol{ax} + a_0$
多元线性回归：
$f(\boldsymbol{x}) = a_1x_1 + a_2x_2 + \cdots + a_nx_n + a_0 = \boldsymbol{ax} + a_0$



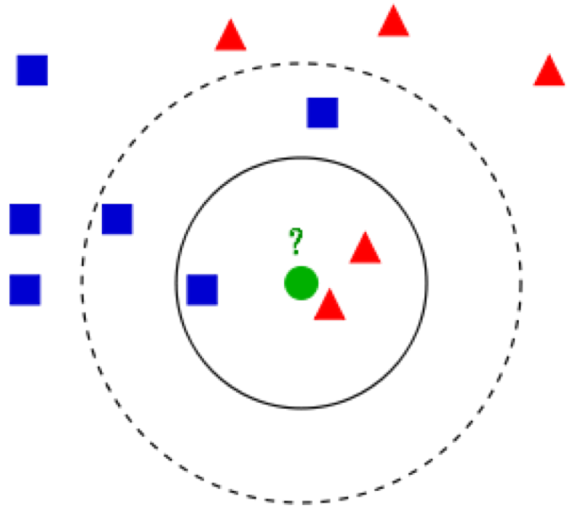
监督学习：分类问题

分类问题——KNN（K最邻近算法）

基本思路：如果一个样本在特征空间中k个最相似的样本中，大多数都是一个类别，那么这个样本也属于这个类别

思路剖析：
定义一个合适的数字k
寻找和待预测样本最“近”的k个样本（在特征空间中最相似）
从这k个样本中找出主要的类别（综合权值最大的类别）

简单粗暴的民主算法



无监督学习：聚类问题

聚类问题——K-Means算法（K-均值）

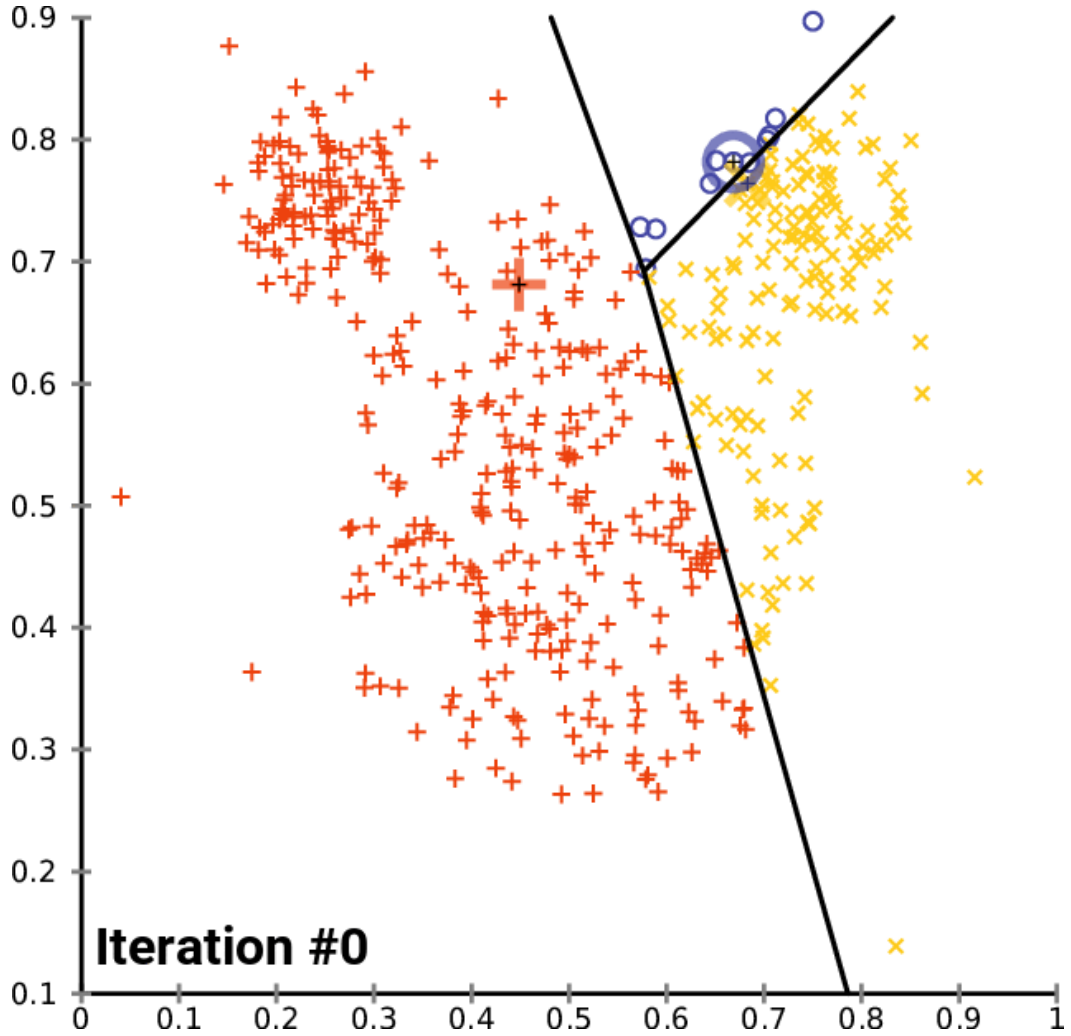
问题描述：
如果有n个未标注的手写数字的图像，如何在不知道分类的情况下将数字先分到10个类别里，每个图像都属于哪一个数字？

问题建模：
需要分析出哪些数字是一类的，不需要知道这个分类到底叫什么

基本思路：
首先假设将数据分成k个分类，不断迭代计算出新的k个分类中心，直到两次迭代的分类中心距离小于一定阈值位置。

K-Means思路剖析

- 假定数据中有k个分类
- 从数据集中选取k个点，作为质心（分类簇的中心点）
- 将每个非质心点关联到距离其最“近”的质心，形成k个簇
- 重新计算k个簇的中心点，作为新的质点，如此往复
- 如果k个中心点的变化趋于“稳定”，那么算法结束，得到最后的质心和所有的聚类结果



小节

机器学习概述

解决哪类问题

如何学习？学习曲线

学习和开发环境准备

使用 scikit-learn以及实验

1.监督学习：

分类问题：花朵分类

回归问题：糖尿病回归分析

2.无监督学习：

聚类问题：手写数字识别数据聚类