统计学习方法读书报告

丁自民

**读书内容：**

1. ，第十二章，第十四章第一节

**读书收获：**

监督学习主要用分类和回归实现，因为数据已经打好标签。直接利用标签信息进行学习即可。

无监督学习主要用聚类和降维。因为数据没有标签，所以无类可分，只能自己找类。

监督学习的效果好一些，同时无监督学习的算法难度又很大。但是，工业界能拿到的数据基本都是没有标签的数据。

**我的疑问：**

P258页 为什么可以用夹角来比较哪几个数据谁更相似？

假如数据点组成的图像大概是2次图像，那么不应该是更贴近这个二次图像的点更加相似吗？

**我的解答：**

答黄一凡第一问：

我猜是人工评估。因为无监督学习有很多参数要自己设（比如某个点是否属于某个类的概率阈值）。所以人们在设置参数的时候，是朝着某个目标去设的。当达到这个目标时，模型就可以了

答郭昊南第一问：

取决于用户需求。

采取聚类时，一般来说用户并不知道自己需要哪些类，聚出来什么就是什么。

采取降维时，用户大概知道要按照什么方向进行降维。

举个例子，给一大推病人进行无监督学习。如果聚类，则可能得到“耳鼻喉病患者”，“幼儿患者”，“住院患者”等等，可能分出许多无厘头的类，这取决于数据的聚集程度。但是如果降维，比如按照年龄进行投影降维，则可能分成“幼儿患者”，“青年患者”，“中年患者”，“老年患者”等和年龄有关的分类。

答殷春锁第三问：

这就是协方差的定义。就像方差定义里面，分母也是n-1一样。