Machine Learning - Week 8

赵燕

目录

Part I
Unsupervised Learning

1 Clustering

聚类算法：非监督学习算法，要让计算机学习无标签数据，而不是之前的标签数据。

1.1 Unsupervised Learning:Introduction

监督学习和无监督学习的比较：

在一个监督学习中，我们有一个有标签的训练集，我们的目标是找到能够区分正样本和负样本的决策边界，我们的监督学习中，有一系列标签，我们需要据此拟合一个假设函数。

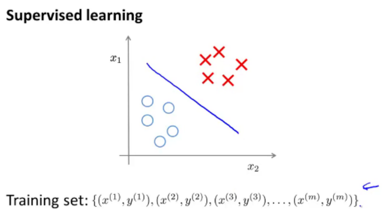


图 1: 监督学习

与此不同的是，在无监督学习中，数据没有附带任何标签，我们拿到的数据就是这样的：

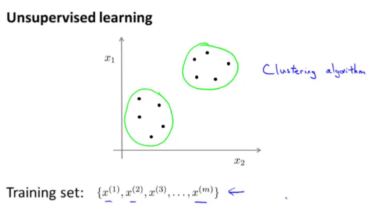


图 2: 无监督学习

在这里我们有一系列的点，却没有任何标签，因此我们的训练集可以写成只有$x_1, x_2, x_3, ..., x_m$,没有任何标签y。因此图上画的这些点没有标签信息，也就是说，在无监督学习中，我们需要将一系列无标签的训练数据，输入到下一个算法中，然后我们告诉这个算法，快去为我们找找这个数据的内在结构给定数据。可能需要某种算法帮助我们寻找一种结构。图上的数据看起来可以分成两个分开的点集（称为簇），一个能够找到我圈出的这些点集的算法，被称为聚类算法。

聚类算法使我们介绍的第一个无监督学习算法，以后还将提到其他的无监督学习算法，它可以为我们找到其他类型的结构或者其他的一些模式，而不是簇。

聚类算法是用来做什么的？

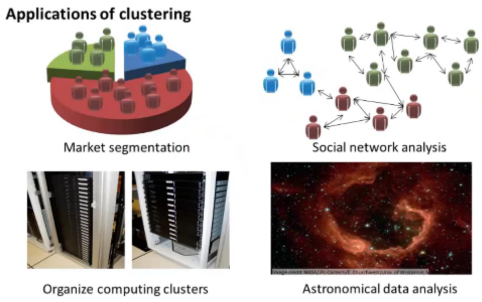
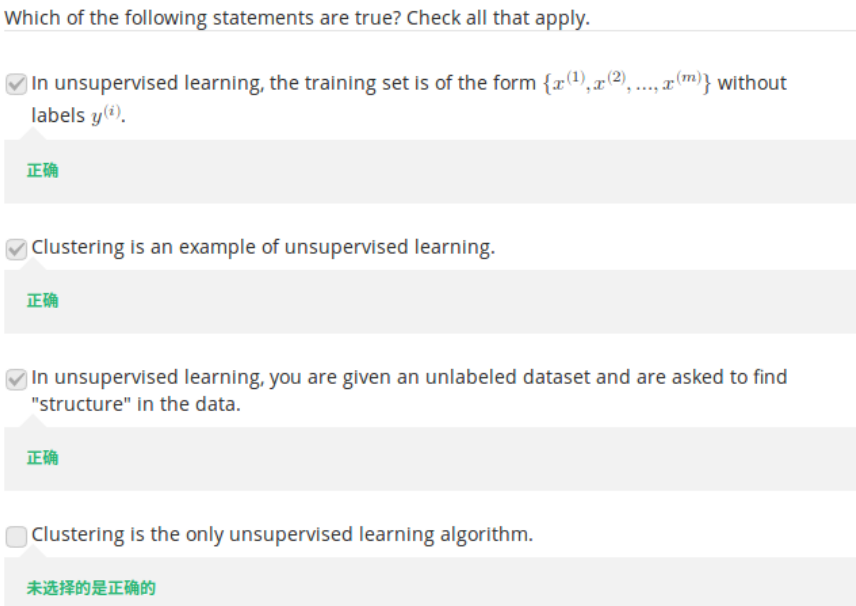


图 3: 聚类算法的应用

- 市场分割:也许你在数据库中分别销售产品或者分别提供更合适的服务，而你希望他们分为不同的课客户群;
- 社交网络分析：例如 Facebook，Google+，或者是其他的一些信息，比如说：你经常跟哪些人联系，而这些人又经常给哪些人发邮件，由此找到关系密切的人群，因此，这可能需要另一个聚类算法，你希望用它发现社交网络中关系密切的朋友。
- 管理数据中心：使用聚类算法来更好的组织计算机集群，或者更好的管理数据中心，因为数据中心中，了解了哪些计算机数据中心更倾向于一起协作工作，那么就可以重新分配资源，重新布局网络，由此优化数据中心，优化数据通信。
- 天文学：利用聚类算法视图了解星系的形成和其中的天文学的具体细节。


Which of the following statements are true? Check all that apply.

In unsupervised learning, the training set is of the form $\{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$ without labels $y^{(i)}$.

正确

Clustering is an example of unsupervised learning.

正确

In unsupervised learning, you are given an unlabeled dataset and are asked to find "structure" in the data.

正确

Clustering is the only unsupervised learning algorithm.

未选择的是正确的

1.2 K-Means Algorithm

K-均值算法是最普及的聚类算法，算法接受一个未标记的数据集，然后将数据聚类成不同的组。

执行K-均值算法，首先先随机选择两个点，这两个点叫做聚类中心，就是图上的两个叉，为什么是两个点呢，因为我们希望聚出两个类。

K-均值是一个迭代方法，需要做两件事情：第一是簇分配，第二是移动聚类中心。

簇分配：需要遍历所有的样本，就是图上绿色的点，然后依据每一个点是更接近红色的这个中心，还是更接近绿色的这个中心，来将每个数据点分配到两个不同的聚类中心中。

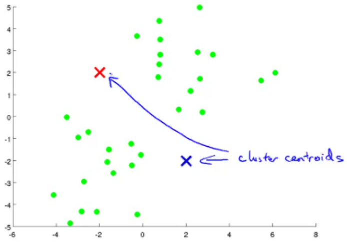


图 4: 簇分配

具体来讲，就是对对数据集中的所有点，根绝他们更接近红色这个中心，还是蓝色这个中心，进行染色，染色的结果如图：

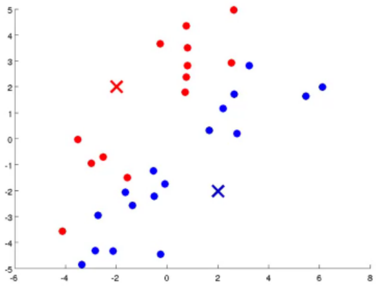


图 5: 簇分配染色

移动聚类中心:将两个聚类中心，也就是说红色的叉和蓝色的叉移动到 和它一样颜色的那堆点的均值处,找出所有红色的点,计算出它们的均值,就是所有红色的点平均下来的位置,然后把红色点的聚类中心移动到这里,蓝色也是这样,找出所有蓝色的点计算它们的均值把蓝色的叉放到那里,我们将按照图上所示这么移动:

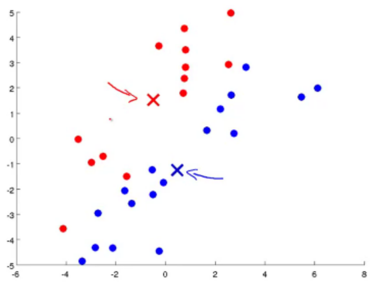


图 6: 移动聚类中心

重复上述步骤，继续簇分配和移动聚类中心，一直迭代下去，直到最后聚类中心不会变，并且哪些点的颜色也不会变，在这时我们就可以说 K-均值方法已经收敛了，在这些数据中找到两个簇。

K-均值是一个迭代算法，假设我们想要将数据聚类成n个组，其方法是：

首先选择K个随机的点，称为聚类中心（cluster centroids）；

对于数据集中的每一个数据，按照距离K个中心点的距离，将其与距离最近的中心点关联起来，与同一个中心点关联的所有点聚成一类。

计算每一个组的平均值，将该组所关联的中心点移动到平均值的位置。

重复步骤直至中心点不再变化。

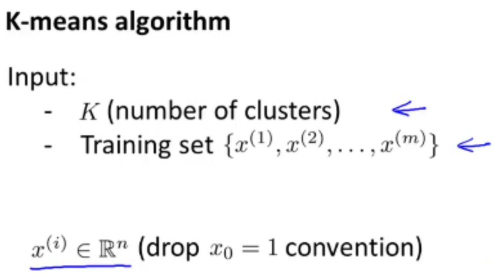


图 7: K-均值算法

$c^{(i)}$ 是距离样本$x^{(i)}$ 的距离的平方，使距离最小或者距离的平方最小都能让我们得到一个相同的$c^{(i)}$，注意$c^{(i)}$ 是一个在1到K之间的数。（通常还是写成距离的平方）

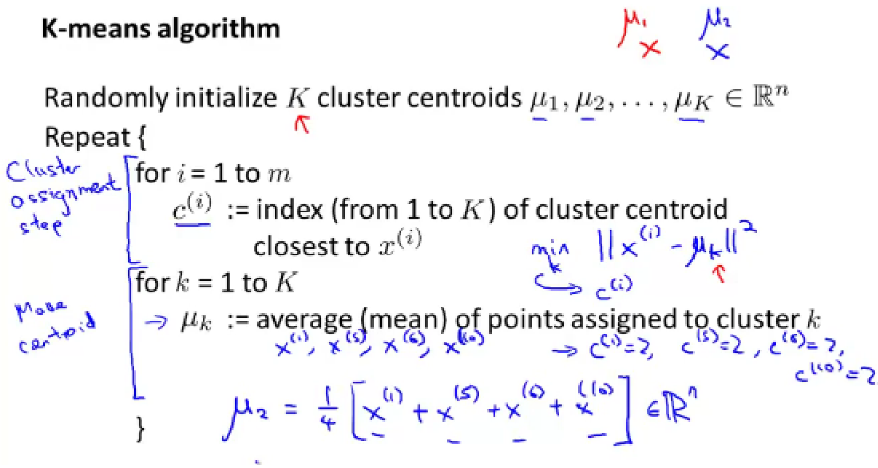


图 8: K-均值算法

算法分为两个步骤,第一个 for 循环是赋值步骤,即:对于每一个样例 i,计算其应该属于的类。第二个 for 循环是聚类中心的移动,即:对于每一个类 k,重新计算该类的聚类中心。

K-均值算法也可以很便利地用于将数据分为许多不同组,即使在没有非常明显区分的组群的情况下也可以。下图所示的数据集包含身高和体重两项特征构成的,利用 K-均值算法将数据分为三类,用于帮助确定将要生产的T-恤衫的三种尺寸 S,M,L。

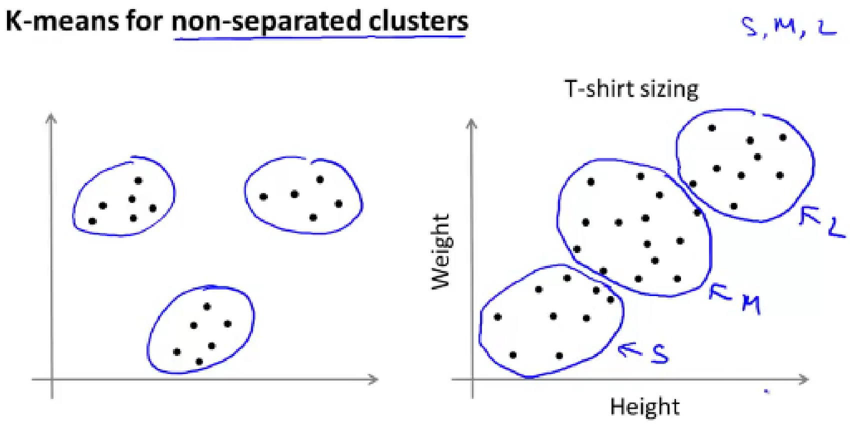


图 9: K-means for non-separated clusters

Suppose you run k-means and after the algorithm converges, you have: $c^{(1)} = 3, c^{(2)} = 3, c^{(3)} = 5, \ldots$

Which of the following statements are true? Check all that apply.

- [x] The third example $x^{(3)}$ has been assigned to cluster 5.

正确

- [x] The first and second training examples $x^{(1)}$ and $x^{(2)}$ have been assigned to the same cluster.

正确

- [ ] The second and third training examples have been assigned to the same cluster.

未选择的是正确的

- [x] Out of all the possible values of $k \in \{1, 2, \ldots, K\}$ the value $k = 3$ minimizes $\|x^{(2)} - \mu_k\|^2$.

正确

1.3 Optimization Objective

1.4 Random Initialization

1.5 Choosing the Number of Clusters

Part II
Dimensionality Reduction

2 Motivation

2.1 Motivation I:Compression

2.2 Motivation II:Visualization

3 Principal Component Analysis

3.1 Principal Component Analysis Problem Formulation

3.2 Principal Component Analysis Algorithm

4 Applying PCA

4.1 Reconstruction from Compressed Representation

4.2 Choosing the Number of Principal Components

4.3 Advice for Applying PCAfg