本周学习：K-均值聚类（K-means）

首先，聚类是一种无监督的学习，它将相似的对象归到同一个簇中。聚类方法可以应用到几乎所有的对象，簇内对象相似度越高聚类效果越好。而K-均值聚类是可以发现K个不同的簇，且每个簇中心采用簇中含值的均值来计算。

簇识别（cluster identification）

簇识别给出聚类结果的含义。聚类与分类的最大不同在于分类的目标事先已知，二聚类不知道。但是它产生的结果与分类相同，只是类别没有事先定义，所以聚类有时又会被称为无监督分类。

聚类就是将相似对象归为统一簇，将不相同对象归为不同簇。相似取决于相似度算法的选择。

K-均值聚类算法

K值由用自己决定。每一个簇通过其质心，即簇的所有点的中心来描述。

算法工作流程：首先，随机确定k个点作为质心。然后将数据集中的每个点分配到每一个簇中，具体来讲，为每个点找距其最近的质心，并将其分配到该质心对应的簇中。这一步完成之后，簇心的值更改为该簇所有点的平均值。

伪代码：

创建k个点作为起始质心（一般是随机选择）

当任意一个点的簇分配结果发生改变时

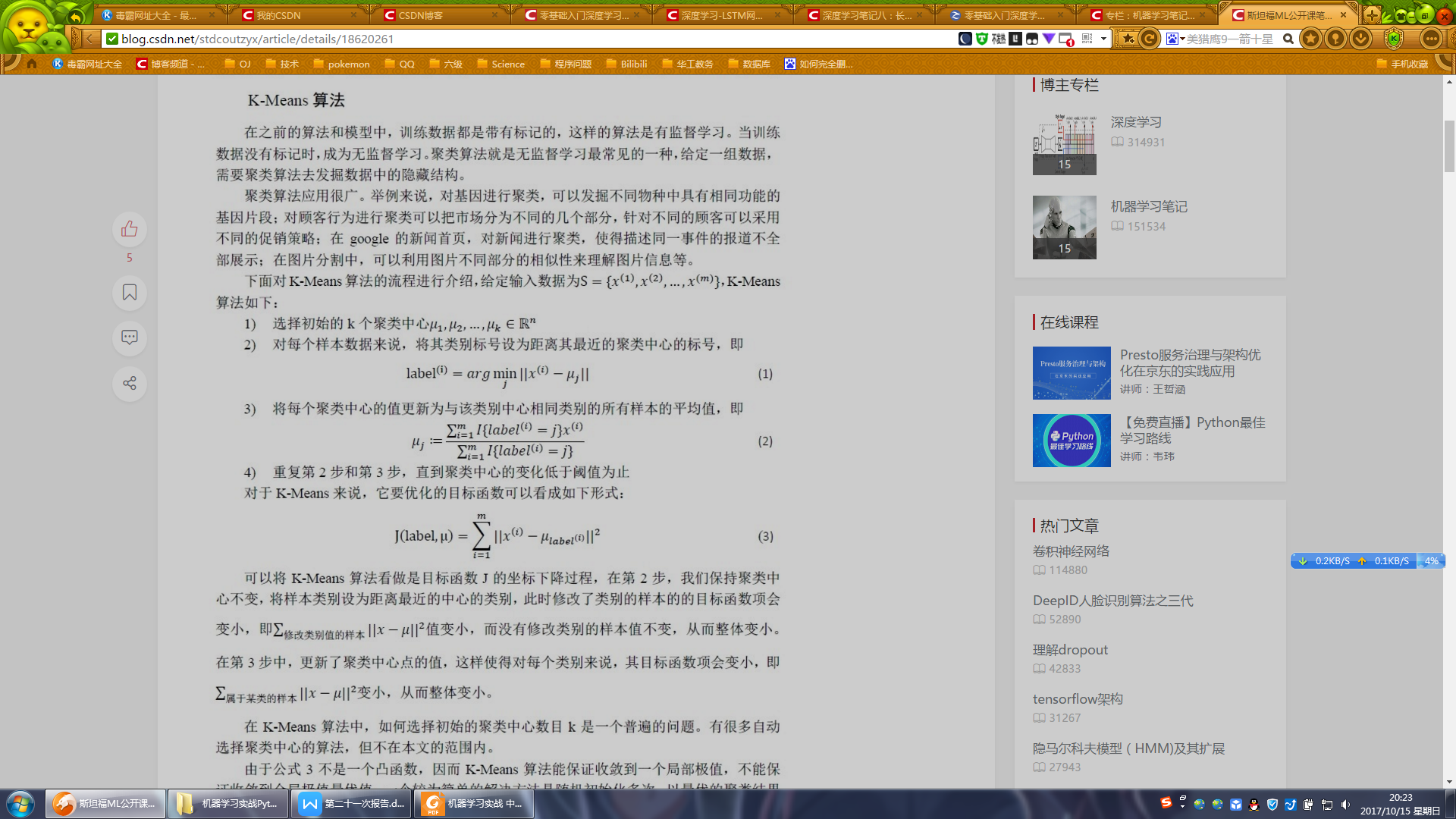
对数据集中的每个数据点

对每个质心

计算质心与数据点之间的距离

将数据点分配到距离它最近的簇

对每一个簇，计算簇中所有点的平均值并将均值作为质心



K-均值聚类一般流程：

1. 收集数据：使用任意方法
2. 准备数据：需要数值型数据来计算距离，也可以将标称型数据映射到二值型数据再用于距离计算
3. 分析数据：任意方法
4. 训练算法：不适用于无监督学习，即无监督学习无训练过程
5. 测试算法：应用聚类算法、观察结果。可以用量化的误差指标来评估算法的结果
6. 使用算法：可以用于所希望的所有应用

评价：

优点：容易实现

缺点：可能收敛到局部最小值，在大规模数据集上收敛较慢

使用数据类型：数值型数据

参考资料：

[1.]机器学习实战-K-均值聚类

[2.]斯坦福ML公开课笔记12——K-Means、混合高斯分布、EM算法