**ML用过的方法的整理**

写在前面：和DL的整理一样，这也是为了高通的intern面试准备的。虽然嘴上说着这辈子再也不搞AI/ML了，但是一个事实就是，未来的尖端科技，必然要与AI/ML相关。虽然目前的职业计划里，我想做软工做系统或者数据库，但是我的经历其实和AI/ML这条路很接近，其实我可以试试。

**OSTU大津**

这是一个非常简单的用来做图像分割的算法，其原理无非就是对于灰度值分为0到m，对于0到m的每一个灰度t，将他作为阈值将图像分割为灰度为0到t以及t + 1到m这两部分。故此，大津可以用作将图像背景和前景分开。

如何找到那个阈值呢？对于每一个阈值，图片分成了两部分，计算每部分所占的比例W1W2和平均灰度值U1U2，总的平均灰度值Ut。计算类间方差（intra-class variance）W1 (U1 - U) \*\* 2 + W2(U2 - U) \*\* 2 = W1W2(U1 - U2)\*\*2。

我们测试了所有阈值之后，简单的选取类间方差时最大的阈值。

**K-mean聚类算法**

**原理**

K-mean就是把数据划成K部分，各个数据找到离自己最近的聚类中心，计算各个数据到聚类中心的距离，聚类中心根据这个距离调整自己的位置，重复直到迭代次数用光或者收敛。K-means 有一个著名的解释：牧师—村民模型：

有四个牧师去郊区布道，一开始牧师们随意选了几个布道点，并且把这几个布道点的情况公告给了郊区所有的村民，于是每个村民到离自己家最近的布道点去听课。

听课之后，大家觉得距离太远了，于是每个牧师统计了一下自己的课上所有的村民的地址，搬到了所有地址的中心地带，并且在海报上更新了自己的布道点的位置。

牧师每一次移动不可能离所有人都更近，有的人发现A牧师移动以后自己还不如去B牧师处听课更近，于是每个村民又去了离自己最近的布道点……

就这样，牧师每个礼拜更新自己的位置，村民根据自己的情况选择布道点，最终稳定了下来。

文本, 信件

描述已自动生成

总体来讲思路是非常简单的，伪代码如下：

文本, 信件

描述已自动生成

**优点**

1. 容易理解，聚类效果不错，虽然是局部最优，但往往局部最优就够了；
2. 处理大数据集的时候，该算法可以保证较好的伸缩性；
3. 当簇近似高斯分布的时候，效果非常不错；
4. 算法复杂度低(O(tknm)，t 为迭代次数，k 为簇的数目，n 为样本点数，m 为样本点维度)

**缺点**

1. K 值需要人为设定，不同 K 值得到的结果不一样，这一点异常显著
2. 对初始的簇中心敏感，不同选取方式会得到不同结果；
3. 对异常值敏感；
4. 样本只能归为一类，不适合多分类任务；
5. 不适合太离散的分类、样本类别不平衡的分类、非凸形状的分类。

**改进**

由于K means聚类算法受到k值的影响非常大，所以有许多方法来找到一个比较优的k值。K means++算法在选择当前的聚类中心时，要选择离之前所有的聚类中心最远的（这样K means maximize每个聚类中的值，才能尽可能的找到全局最优），并且以一定的概率选择新的中心点。