第十章作业

陈镜融
14307130118@fudan.edu.cn

May 8, 2018

1 10.4

For the k-means algorithm, it is interesting to note that by choosing the initial cluster centers carefully, we may be able to not only speed up the algorithm's convergence, but also guarantee the quality of the final clustering. The k-means++ algorithm is a variant of k-means, which chooses the initial centers as follows. First, it selects one center uniformly at random from the objects in the data set. Iteratively, for each object p other than the chosen center, it chooses an object as the new center. This object is chosen at random with probability proportional to dist(p)2, where dist(p) is the distance from p to the closest center that has already been chosen. The iteration continues until k centers are selected. Explain why this method will not only speed up the convergence of the k-means algorithm, but also guarantee the quality of the final clustering results.

k-means算法存在如下问题

1 k需要事先指定

2 k-means在开始聚类之前，需要事先选好k个聚类中心

3 k-means对离群值非常敏感，聚类结果准确性受边缘数据影响很大

对第一点，可以选择一个评价指标，之后对k进行枚举，来作为workaround。 对第三点，可以采用k-Medoids的方法进行改进。 对第二点，初始聚类中心的选择，对收敛速度和最终结果都有很大影响，因此Arthur等人在2007年提出了k-means++的算法[1]。

该算法主要对k-means初始聚类中心的选择作了改进。对于普通k-means算法，我们在样本空间中随机选择k个点作为初始聚类中心，而k-means++，在选第i个初始点的时候，会以与当前最近点的欧几里得距离平方成正比的概率，选择这次的聚类中心，也就是说，距离当前空间聚类中心越远，就有更大的概率被选取。

这使得初始点更加分散，从直观上理解，这样的初始状态也更加接近我们希望最后聚类出来的结果。因此我们期望k-means++能在收敛速度上和训练质量上有更好的结果。

根据论文[1]，k-means从来都不是因为其准确性，而是因为其训练速度收到青睐的。文章大部分笔墨用来介绍初始点的选取方法和对potential function满足的性质的一些证明。

论文中对训练收敛速度的提升，是通过Empirical Results一章节中，利用实际训练结果来作说明的，并没有给出理论证明。

	Average ϕ		Minimum ϕ		Average T	
k	`k-means`	`k-means++`	`k-means`	`k-means++`	`k-means`	`k-means++`
10	135512	126433	119201	111611	0.14	0.13
25	48050.5	15.8313	25734.6	15.8313	1.69	0.26
50	5466.02	14.76	14.79	14.73	3.79	4.21

Table 2: Experimental results on the *Norm-25* dataset (n = 10000, d = 15)

	Average ϕ		Minimum ϕ		Average T	
k	`k-means`	`k-means++`	`k-means`	`k-means++`	`k-means`	`k-means++`
10	7553.5	6151.2	6139.45	5631.99	0.12	0.05
25	3626.1	2064.9	2568.2	1988.76	0.19	0.09
50	2004.2	1133.7	1344	1088	0.27	0.17

Table 3: Experimental results on the *Cloud* dataset (n = 1024, d = 10)

	Average ϕ		Minimum ϕ		Average T	
k	`k-means`	`k-means++`	`k-means`	`k-means++`	`k-means`	`k-means++`
10	$3.45 \cdot 10^8$	$2.31 \cdot 10^7$	$3.25 \cdot 10^8$	$1.79 \cdot 10^7$	107.5	64.04
25	$3.15 \cdot 10^8$	$2.53 \cdot 10^6$	$3.1 \cdot 10^8$	$2.06 \cdot 10^6$	421.5	313.65
50	$3.08 \cdot 10^8$	$4.67 \cdot 10^5$	$3.08 \cdot 10^8$	$3.98 \cdot 10^5$	766.2	282.9

Table 4: Experimental results on the *Intrusion* dataset (n = 494019, d = 35)

Figure 1:

实验结果显示了当n增大时，训练速度将提升明显，而没有提到其准确性的提升。因此最终聚类结果的质量提升，我觉得是不一定能保证的。而速度的提升，也是由实验得出的结果。其原因，我觉得主要在于，训练刚开始阶段，聚类中心在往最终目标移动的速度会比较缓慢，而k-means++一开始就让种子更加接近最后结果，因此省去了这部分训练的时间，因此能有速度上的提升。

Reference

[1] Arthur, D., and Vassilvitskii, S. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (2007), Society for Industrial and Applied Mathematics, pp. 1027–1035.