67750: Advanced Practical Machine Learning	2019/20

Exercise 4 — Clustering Theoretic Solutions

Dr. Tommy Kaplan — *TA: Omri Bloch*

1 Theoretical Questions

1.1 Convergence of k-means

Prove that k-means converges in a finite number of steps. (Hint: what happens to the cost function $\sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$ at each iteration?)

Solution

Observe that there are a finite amount of possible clusterings that the algorithm can possibly go through, which correspond to all the possible ways of dividing n data points into k clusters - k^n.

Next, we will show that the cost function, $\sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$, is monotonically decreasing with each step of the algorithm:

1. In the centroid step, the chosen centroids of each cluster are chosen as to minimize the cost of their own cluster, $\sum_{x \in C} ||x - \mu||^2$. This means that the cost function can only decrease in this step.
2. In the clustering step, each point is allocated to the centroid that is closest to it, such that $C(x) = argmin_k(||x - \mu_k||^2)$. This means that the cost function can only decrease in this step too.

Since there are a finite amount of clusterings for our problem, and we only decrease our cost function, we won't return to the same clustering twice. This means that eventually we will stay in the same clustering, which means we have converged.

1.2 Suboptimality of k-means

1. Give an example of data points and an initialization of k-means which converge to a suboptimal solution (a local minima of the cost function).
2. Give an example of data points that have more than one optimal clustering solution (with respect to the cost function).

Solution

1. arranging 4 points in a rectangle: $\{(0,0),(0,1),(2,0),(2,1)\}$ and initializing with the two centroids $\{(1,0),(1,1)\}$ would be a suboptimal solution (the optimal being the centroids $\{(0,0.5),(2,0.5)\}$).
2. arranging 4 points in a square: $\{(0,0),(0,1),(1,0),(1,1)\}$ has two optimal clusterings: $C_1 = \{(0,0),(0,1)\}$ $C_2 = \{(1,0),(1,1)\}$ and $C_1 = \{(0,0),(1,0)\}$ $C_2 = \{(0,1),(1,1)\}$.

1.3 Initializations of k-means

Suppose we want to find a globally optimal solution of a k-means clustering problem. We'll assume that if the algorithm is initialized such that every cluster is represented in the random starting points then the algorithm will converge to the optimal solution. That is, a "good" event is one where each example in the initial guess is from a different cluster. Denote the probabilities of sampling a point from each cluster by $\alpha_1, ...\alpha_k$, and:

1. Calculate the expected number of trials we have to perform before we sample a good initialization event (as a function of k, α_i). Do this by calculating the probability of a single "good event", then use the fact that the number of trials is a geometric random variable to get the expectation easily.
2. Give a lower bound on this number as a function of k. Do this by finding the best possible α variables, subject to $\sum_i \alpha_i = 1$.
3. Is this a plausible number of start trials? If not, how can we explain the fact that K-Means often works?

Solution

1. Assuming a uniform point selection, the expectation can be written as the expectation of a geometric random variable with:
$$p = k! \prod \alpha_i \implies \mathbb{E}[\#\text{trials}] = \frac{1}{k! \prod \alpha_i}$$
2. A lower bound independent of the α's:
$$\mathbb{E}[\#\text{trials}] = \frac{1}{k! \prod \alpha_i} \geq \min_{\alpha_i} \frac{1}{k! \prod \alpha_i} = \frac{1}{k! \max_{\alpha_i} \prod \alpha_i}$$
It's left to find the maximum of the function $\prod_i \alpha_i$ subject to $\sum_i \alpha_i = 1$. Intuitively, the optimum is obtained when $\alpha_i = \alpha_j = \frac{1}{k}$ for all i, j. To show that this is the true, we'll use Lagrange multipliers and the log trick:
$$arg \max_{\alpha} \prod \alpha_i = arg \max_{\alpha} log(\prod \alpha_i) = arg \max_{\alpha} (\sum_{i=1}^{k} log(\alpha_i))$$
$$L(\alpha, \lambda) = \sum_{i=1}^{k} log(\alpha_i) - \lambda(\sum_{i=1}^{k} \alpha_i - 1)$$
We will differentiate w.r.t. an arbitrary α:
$$\frac{\partial L}{\partial \alpha_i} = \frac{1}{\alpha_i} - \lambda = 0 \rightarrow \alpha_i = \frac{1}{\lambda}$$
And we see that all of the α's equal the same value, so the answer (upholding our constraint) is as we expected, $\forall \alpha : \alpha_i = \frac{1}{k}$.
So we have that:
$$\mathbb{E}[\#\text{trials}] \geq \frac{1}{k! \max_{\alpha_i} \prod \alpha_i} = \frac{k^k}{k!}$$
And using, e.g., Stirling's approximation, it's evident that the expected number of trials exponential in k.

3. This is obviously a bad result, and it's probably a very conservative lower bound (the result is much worse if the α_i's are very different). How do we cope? first - it's not clear that the assumption about the initialization distribution is required - it's easy to devise an example (say with $k = 2$), where the initialization point is irrelevant to the final result, which is optimal in any case. So maybe this assumption could be relaxed. Secondly - we can simply select the starting point in a more intelligent way (e.g. kmeans++).