Lecture 3 Notes

Scribe: Lisa Anne Hendricks

Presenters: Nick Boyd and Nick Altieri

This weeks theme was importance sampling. We looked at how it can be applied to k-means ++ and universal ϵ-approximations for integrals.

1 k-means++: The Advantages of Careful Seeding

The goal is to represent a large set of data by k clusters. Begin with the potential function:

$$\phi(z, c_1^k) = \sum_{i=1}^{n} \sum_{j=1}^{k} 1(z_i = j)\|x_i - c_j\|_2^2$$

$$\text{where } z \in \{1, 2, ...k\}^n \text{ and } c_i \in R^n$$

The data is summarized with k vectors. Each point x_i in the set is represented as one of the k vectors $\{c_1, ..., c_k\}$. The loss for doing this is $\|x_i - c_j\|_2^2$. z are the optimal clusters and c are the centers. Note the duality between z and c: given the centers, the decision boundaries for optimal clusters can be found and given the clusters, the optimal center can be found. Thus we can write $\phi(z)$. This problem is non-convex and NP hard to solve. A common method to determine centers is Lloyd's method described in the next section.

1.1 Lloyd's Method

Lloyd's method is as follows:

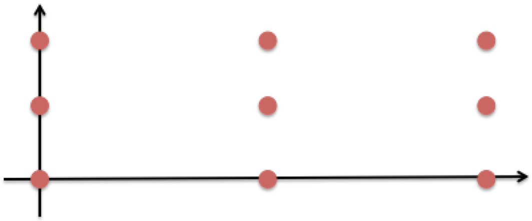
1. Choose k initial centers by uniformly sampling from all points.

2. For each center c_i, set the corresponding cluster z_i to be the set of points such that all points in z_i are closest to c_i.

3. Set the new centers to be the center of mass of the clusters.

4. Repeat steps 2 and 3.

Lloyd's method works okay in practice, but there are many sets of points we can construct where Lloyd's method will likely fail. For example, consider the set of data below:



In step 1, it is likely that centers are not chosen from each cluster, and Lloyd's method will yield results far from optimal. The k-means++ algorithm addresses this by introducing a different sampling method.

1.2 k-means++

Define $D_c(x) = \min_j \|x_i - x_j\|_2$ and note that $\phi(c) = \sum_{i=1}^{n} D_c^2(x_i)$. Now choose one center, c_1 uniformly from the set, then choose the next center such that $P(c_j = x_i) \propto D^2(x_i)$. Once these centers have been determined, Lloyd's method can be used. Sampling using the above scheme means that once a center c_i is picked, points further away from this center and thus not in the corresponding cluster are more likely to be picked. In short, the method aims to select starting centers in each cluster. At this point, we asked what D_c would look like if we thought of the objective function as a Guassian mixture probability. The result is:

$$\max_z(log(p(x_i, z_i|c) \propto -D_c^2(x_i))$$

The expected value of the objective function determined using this new sampling method has the following bound: $E[\phi(c)] \leq 8{*}ln(k+2)\phi(c{*})$ where c* is the optimal clustering. To understand the bounds on the expected value, first consider the contribution to the cost of a set of points A in the set:

$$\phi(c, A) = \sum_{x \in A} \min_j \|c_j - x\|_2^2$$

If A consists of all the points, then the above is just our objective function. Now consider:

$$\begin{gathered} E[\phi(c_i, c_1{*})|c_i \in c_1{*}] \\ = E[\sum_{x \in C_1{*}} \|c_j - x\|_2^2 | c_i \in c_1{*}] \end{gathered}$$

$$= \sum_{c_i \in C_1*} \frac{1}{|c_1*|} \sum_{x \in C_1*} \|c_j - x\|_2^2$$

Before continuing, note the following lemma:

Consider a set of points $S \in x$:

$$\sum_{x \in S} = \|x - z\|_2^2 = \sum_{x \in S} \|x - \bar{s}\|_2^2 + |c| \|\bar{s} - z\|_2^2$$

This can be proven by expanding out the inner products. Using this:

$$\sum_{c_1 \in C_1*} \frac{1}{|c_1*|} [\phi(c*, c_1*) + |c_1|^2 \|c_1 - c*\|_2^2]$$

$$= 2\phi(c*, c_1*)$$

Now consider we have chosen c centers and choose the next center,c'. It can be shown that:

$$\phi([c, c'], C*) | c' \in C*] \leq 8\phi(c*, C*)$$
$$= \sum_{x \in C} \min(D_c(x), \|x - c'\|_2)^2 | c' \in C*]$$

The minimum in the sum means that the cost is less if x is closer to the added point c'.

$$= \sum_{c' \in C*} \frac{D_c^2(c')}{\sum_z D_c^2(x)} \sum x \in C* \min(D_c(x), \|x - c'\|_2)^2$$

In order to continue, we need to bound $D_c^2(c')$.

$$D_c(c') = \min_i \|c_i - c'| \ \leq \min(\|c_1 - x\| + \|x - c'\| (\text{for all x}) = D_c(x) + \|x - c'\|$$

$D_c^2(c')$ can be bounded using Jensen's inequality which leads to:

$$D_c^2(c') \leq 2D(x)^2 + 2\|x - c'\|^2$$

Averaging over $x \in C^*$ leads to

$$D_c^2(c') \leq \frac{2}{|c^*|}[\sum_{x \in C^*} D(x)^2 + \|x - c'\|^2]$$

Plugging this back in, everything cancels and works!

The sampling method increases the chances that a point will be picked in every cluster. If a point in every cluster is chosen, the solution is competitive with the optimal solution.

2 Universal epsilon-approximations for integrals

Consider the means function define $f_{U_1}, ...f_{U_k} = \min_i \|x - u_i\|^\alpha$. This function class is F_α. We should like to choose a set of points $R \in S$ such that for all f in F, $\sum\limits_{x \in R} f(x)\upsilon(x) \in (1 \pm \epsilon) \sum x \in Sf(x)$. Professor Recht has discussed unreasonably effective algorithms and Nick (the presenter) argued that there are some algorithms that are offensively effective, such as uniformly sampling. Thus, before we continue we should consider if uniformly sampling will work. However, for similar reasons discussed in the previous paper, uniformly sampling will not necessarily work well, so we need to do something smarter.

Note that for any distribution function:

$$E[f] = \frac{1}{n} \sum_{x \in S} f(x) = \sum_{x \in S} \frac{f(x)}{nq(x)} q(x)$$

If we sample via $q(x)$ and compute $T_f(x) = \frac{f(x)}{nq(x)}$, $T_f(x)$ will be an unbiased estimator of $E[f]$. However, we must also consider the variance.

Define the sensitivity of a function be defines as $\sigma(f)$ as $\sup_{f \in F} \frac{f(x)}{f}$.

At this point, we went on a short aside about taking the sup of a function. Professor Recht had us consider evaluating the following:

$$\sup_x \frac{(a_i x - b_1)^2}{\sum\limits_{i=1}^{n} (a_i x - b_1)^2}$$

where $\|Ax - b\|$ is an overdetermined system. The final result is just the leverage scores $\|V^T e_i\|^2$ where the $US^2V = \sum\limits_{i=1}^{n} a_i a_i^T$. Intuitively, taking the supremum of a function will keep around the big terms. Hence, the sensitivity looks at how much each x contributes to the objective function.

Let the total sensitivity of a function class be $\mathfrak{S} = \sum\limits_{x\in S} \frac{1}{n}\sigma_F(x)$. Let $s_F(x) \geq \sigma_F(x)$ and $S(F) = \sum\limits_{x\in S} \frac{1}{n} s_F(x)$.

Pick $q(x)$ sensitivity such that $q(x) = \frac{S_F(x)}{nS(F)}$.

After choosing $q(x)$, we can evaluate the variance.

$$\text{Theorem: } Var(T_f) \leq (S(F)-1)\bar{f}^2$$

$$\begin{aligned}\text{Proof: } \frac{1}{\bar{f}^2} Var(T_f) &= \frac{1}{\bar{f}^2} \sum x \in S(\frac{f(x)}{nq(x)} - \bar{f})^2 q(x) \\ &= \frac{1}{\bar{f}^2} \sum_{x\in S} (\frac{s_F(x)}{nS(F)}) * (\frac{f(x)S(F)}{s_F(x)}) - \bar{f}))^2.\end{aligned}$$

Expanding and simplifying this becomes:

$$\frac{1}{\bar{f}^2} \sum_{x\in S} (\frac{f(x)^2 S(F)}{n s_F(x)}) - 2 + 1$$

Note that for any $f \in F$ and $x \in X$, $f(x) \leq s_F(x)\bar{f}$. Therefore,

$$\frac{1}{\bar{f}}^2 \sum_{x\in S} \frac{1}{n} f(x)\bar{f}S(F) - 1 = S(F) - 1$$

If we choose a random sample R w.r.t distribution q(x) with $|R| = a * (\frac{2(S-1)}{\epsilon^2})$, then

$$Pr[|\bar{f} - \frac{1}{a} \sum_{x\in R} (\frac{S(F)f(x)}{s(x)}| \geq \epsilon \bar{f}] \leq \frac{1}{2}$$

Proof:

Recall the Chebyshev inequality: $P[|x - \mu| \geq k\sigma] \leq \frac{1}{k^2}$

$$\sigma = \sqrt{Var(\frac{1}{a}\sum_{x\in R} T(x))} \leq \sqrt{\frac{1}{a}(S(F)-1)\bar{f}^2}$$
$$= \sqrt{\frac{\epsilon^2(S(F)-1)\bar{f}^2}{2(S(F)-1)}} = \frac{\epsilon\bar{f}}{\sqrt{2}}$$

$$Pr[|\bar{f} - \frac{1}{a}\sum_{x\in R}(\frac{S(F)f(x)}{s(x)}| \geq \epsilon\bar{f}] \leq \frac{1}{2}$$

After this proof, there was some discussion on using Chebyshev as a bound and if a Hoeffding inequality could be used as a better bound.

3 Conclusion

In this lecture, we saw how importance sampling can perform better than uniform sampling in k-means and approximation of integrals. The underlying principle in each paper is that samples that contribute more, or are of more importance, should be more heavily weighted.