Stat 991: Multivariate Analysis, Dimensionality Reduction, and Spectral Methods **Lecture: 9**

Clustering; Single Linkage; and Pairwise Distance Concentration

Instructor: Sham Kakade

1 For a Gaussian

Assume that $x \sim \mathcal{N}(0, \Sigma)$. Now if we project to the k-dimensional subspace corresponding to the span of the top k singular vectors, then this corresponds to preserving the k dimensional subspace with maximal variance.

But what about multi-modal distributions? The next few lectures will try to provide some intuition as to the outcomes of various projections of mixtures of Gaussians.

2 Isotropic Gaussians and Intuition

Now let us consider the case where $X \in \mathbb{R}^n$ is distributed according to k-mixture of Gaussians. In particular, assume each component Gaussian is isotropic, e.g. $\mathcal{N}(\mu_i, \sigma_i^2 \mathrm{I})$ for the i-th component. Let π denote the mixing weights.

Also, let us assume the means $\{\mu_i\}$ are non-degenerate, e.g. there is a unique k-dimensional subspace containing these k points.

2.1 In Low Dimensions

In low dimensions, e.g. when n is "small", and if the Gaussians are separated, then we expect a high density region to contain the mean. So we can simply just find regions of points which are close together, and a natural estimate of the cluster mean is the average of some tightly grouped set of points.

2.2 The Chi-Square tail bound

Recall Lemma 1.3,our tail bound for χ^2 variables with n degrees of freedom, which stated that: $\Pr(\chi_n^2 \leq (1+\epsilon)n) \leq \exp(-\frac{n}{4}(\epsilon^2 - \epsilon^3))$ (there is also a lower bound). For the case of $eps < 1/2$, we have that:

$$\Pr(|\chi_n^2 - n| \leq \epsilon n) \leq 2\exp(-\frac{n}{8}\epsilon^2)$$

where the factor of 2 is since we are using both the upper and lower tail bounds. In other words:

$$\Pr(|\chi_n^2 - n| \leq \epsilon\sqrt{n}) \leq 2\exp(-\frac{\epsilon^2}{8})$$

This provides us with the intuition for the high dimensional case.

2.3 In High Dimensions

But what about high dimensions? What is the density of the points near the mean? And how far away is the average point from it's component mean? Let us address this questions for a single isotropic Gaussian distribution. First, note that $\mathbb{E}[\|x\|^2] = n\sigma^2$. Hence, on average, we expect a point to be rather far from mean, but let us quantify this. Recall, that the distribution of $\|x\|^2$ is a χ_n^2 distribution with n degrees of freedom. Hence, the variance of $\|x\|^2$ is $2n\sigma^4$ (so the deviation if $\sqrt{2n}\sigma^2$). Hence, we expect the average distance to the mean to be $\sqrt{n}\sigma \pm O(n^{1/4}\sigma)$; so not only do we expect the points to be far away, they also will be quite far away with low variance.

In particular, how many samples to we need in order to have a point $\frac{1}{2}\sqrt{n}\sigma$ away from the mean? Our chi square tail bound from above implies that we need $\Omega(2^n)$ samples to get a point merely $\frac{1}{2}\sqrt{n}\sigma$. Note for this case we need to we need to set $\epsilon = O(\sqrt{n})$.

This rather gloomy observation motivates the use of dimensionality reduction techniques for clustering.

2.4 Aside: "every point is an outlier"

Often one hears the colloquial expression, "in high dimensions, every point is an outlier". Let us understand this remark, in light of the previous discussion.

3 Separation and Distance Based Clustering

Now let us consider perhaps the most simplest case, where all the intraclass distance between points are smaller then the interclass distances. More precisely, let us suppose we have a dataset in which, for all x and x' generated from the same cluster, we have that $\|x - x'\| \leq \|x - x''\|$ where x'' was generated from another cluster. In this setting, what clustering algorithms succeed?

3.1 Single Linkage Clustering; Minimum Spanning Trees; and Hierarchical Clustering

From Wikipedia:

In cluster analysis, single linkage or nearest neighbor is a method of calculating distances between clusters in hierarchical clustering. In single linkage, the distance between two clusters is computed as the distance between the two closest elements in the two clusters. Mathematically, the linkage function — the distance D(X,Y) between clusters X and Y — is described by the following expression : D(X,Y) = min(d(x,y)) where d(x,y) is the distance between elements and ; X and Y are two sets of elements (clusters) A drawback of this method is the so-called chaining phenomenon: clusters may be forced together due to single elements being close to each other, even though many of the elements in each cluster may be very distant to each other. [edit]Algorithm

The following algorithm is an agglomerative scheme that erases rows and columns in a proximity matrix as old clusters are merged into new ones. The proximity matrix D contains all distances d(i,j). The clusterings are assigned sequence numbers 0,1,......, $(n-1)$ and L(k) is the level of the k-th clustering. A cluster with sequence number m is denoted (m) and the proximity between clusters (r) and (s) is denoted d[(r),(s)]. The algorithm is composed of the following steps:

1. Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.

2. Find the least dissimilar pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.

3. Increment the sequence number: m = m + 1. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)]

4. Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined as d[(k), (r,s)] = min d[(k),(r)], d[(k),(s)].

5. If all objects are in one cluster, stop. Else, go to step 2.

Now, when will single linkage clustering "succeed"?

Lemma 3.1. *Assume that all intraclass distances are smaller than interclass distances. Then the single linkage algorithm will never merge points from two different clusters, if at all possible. Hence, if the algorithm is stopped when there are precisely k clusters, then all points belong in each component belong to the same cluster.*

Proof. fill me in □

3.2 "Naively" Separated Cluster

Let us understand when this occurs. Let us denote the $\overline{\mu}_i$ denote the sample mean and let $\overline{r}_i$ denote the radius of the points in cluster i. Then if:

$$\|\mu_i - \mu_j\| \geq r \max\{\overline{r}_i, \overline{r}_j\}$$

Then all intraclass distances are smaller than interclass distances.

Recall that with very high probability, we will have that each point is close to mean within $2\sqrt{n}\sigma_i$ (in the isotropic case). In fact, just with a union bound/Bonferoni argument, we have that all points m points (for a dataset of size m) are $2\sqrt{n}\sigma_i\sqrt{\log m/\delta}$ close to their respect means, with probability greater than $1-\delta$.

Hence, a sufficient condition for single linkage to succeed, with high probability, is for:

$$\|\mu_i - \mu_j\| \geq C \max\{\sigma_i, \sigma_j\}\sqrt{n \log m/\delta}$$

where C is a constant.

Note that this previous argument is rather stringent in that we are demanding that the entire Euclidean balls (which contain each cluster) be completely separated. Can we do better?

3.3 More Subtly...

In essence, the question is one of: how far apart do we need to push two spheres apart such that (uniformly) random points from the sphere tend to be closer than random points from the two different spheres?

Let X_i and X_j be samples from clusters i and j. Let us assume that allow isotropic Gaussians have the same variance, for simplicity (e.g. $\sigma_i = \sigma_j$). We have that:

$$X - Y \overset{d}{=} N(\mu_i - \mu_j, 2\sigma^2 I_d) \overset{d}{=} \mu_i - \mu_j + \sqrt{2}\sigma W$$

where W is a standard normal. Hence,

$$\|X_i - X_j\|^2 = \|\mu_i - \mu_j\|^2 + 2\sigma\|W\|^2 + 2\sqrt{2}\sigma(\mu_i - \mu_j) \cdot W$$

For the last term, we have:

$$(\mu_i - \mu_j) \cdot W \overset{d}{=} (\mu_i - \mu_j) \cdot Z$$

where Z is a standard normal — this is due to that we are considering the variance in one direction (the $\mu_i - \mu_j$ direction). Intuitively, this second term is very near 0 and negligible in our argument. To deal with it formally, we use that $2a \cdot b \leq \|a\|^2 + \|b\|^2$, so that:

$$\|X_i - X_j\|^2 \geq 2\|\mu_i - \mu_j\|^2 + 2\sigma^2\|W\|^2 - 2\sigma Z^2$$

Similarly, for X_i and X_i' sampled from the same component cluster:

$$\|X_i - X_i'\|^2 \quad = \quad 2\sigma^2\|W\|^2$$

Hence, by our tail bounds for χ_n^2 distributions, we know that, for all m samples, $|\|W\|^2 - n| \leq \sqrt{n\log(m/\delta)}$ (with probability greater than $1-\delta$). Similarly Z^2 is a χ^2 with one degree of freedom, so that $|Z^2 - 1| \leq \sqrt{\log(m/\delta)}$. Hence:

$$\|X_i - X_j\|^2 \geq 2\|\mu_i - \mu_j\|^2 + 2\sigma^2 n - 2\sigma^2 C\sqrt{n\log(m/\delta)}$$

where C is a constant. Similarly:

$$\|X_i - X_i'\|^2 \leq 2\sigma^2 n + 2\sigma\sqrt{n\log(m/\delta)}$$

(Caveat: we need to use union bound, since three are three failure events, each with probability δ. But this just means changing $\delta \to \delta/3$).

Hence, for single linkage to succeed, we desire that:

$$\|\mu_i - \mu_j\| \geq C\sigma\left(n \log \frac{m}{\delta}\right)^{1/4}$$

For the case of non-equal variances (but still isotropic), we can replace σ with $\max\{\sigma_i, \sigma_j\}$ (with some minor technical assumptions, which essentially say that one Gaussian can not essentially live in another Gaussian).

4 "Minimal Separation" for well separated clusters

It is often natural to only consider clustering problem when the data are "well separated" — one viewpoint is that clustering is only interesting when the data are in fact well "separated". However, we can ask the question of what is the separation under which, with high probability, each point can be correctly identified with the appropriate generating cluster, e.g. the Bayes optimal classification (with knowledge of the generative parameters) has a hight probability of success.

Here, it it turns out that with:

$$\|\mu_i - \mu_j\| \geq C \max\{\sigma_i, \sigma_j\} \left(\log \frac{m}{\delta}\right)^{1/2}$$

then all points will, with probability greater than $1 - \delta$, be classified correctly. Importantly, note this lower separation has no n dependence.

Hence, there is a dramatic gap in what we are able to cluster (under single linkage) in comparison to what might be considered a lower bound. Currently, there are essentially no known computationally better estimators.