[Lecture 23]

1 Agnostic Clustering

We next study the idea of partial specification for clustering. Our setting for clustering will be the following:

- There are k unknown distributions $p_1, \ldots, p_k$.
- We observe points $x_1, \ldots, x_n$, such that a fraction α_j of the points x_i are drawn from p_j.

Generally the α_j are not known but we have a lower bound on $\alpha_{\min} = \min_{j=1}^{k} \alpha_j$. In clustering we have two goals:

- **Parameter recovery:** We wish to estimate some parameter of the p_j (usually their means).
- **Cluster recovery:** We wish to determine for each point x_i which cluster p_j it was drawn from.

In the simplest setting, we assume that each of the p_j has a known parametric form (for instance, each p_j is a Gaussian with unknown mean and variance). In the *agnostic* setting, we do not assume a parametric form for the p_j but instead only assume e.g. bounded moments. In the *robust* setting, we allow some fraction ϵ of the points to be arbitrary outliers (so $\alpha_1 + \cdots + \alpha_k = 1 - \epsilon$).

Partial specification thus corresponds to the agnostic setting. Clustering is a particularly interesting setting for studying partial specification because some algorithms that work in the simple setting fail completely in the agnostic setting. Below we will first study the simple setting and give an algorithm based on the method of moments, then turn our attention to the agnostic setting. In the agnostic setting, resilience will appear once again as an information-theoretically sufficient condition enabling clustering. Finally, we will turn our attention to efficient algorithms. In many cases the agnostic algorithms will work even in the robust agnostic setting.

1.1 Clustering Mixtures of Gaussians

Here we assume that each $p_j = \mathcal{N}(\mu_j, \Sigma_j)$. Thus we can treat each x_i as being drawn from $p = \sum_{j=1}^{k} \alpha_j \mathcal{N}(\mu_j, \Sigma_j)$. This is a parametric model with parameters $(\alpha_j, \mu_j, \Sigma_j)$, so (at least in the limit of infinite data) a sufficient condition for exact parameter recovery is for the model to be identifiable, meaning that if $\sum_{j=1}^{k} \alpha_j \mathcal{N}(\mu_j, \Sigma_j) = \sum_{j=1}^{k} \alpha'_j \mathcal{N}(\mu'_j, \Sigma'_j)$, then $\alpha_j = \alpha'_j$, $\mu_j = \mu'_j$, and $\Sigma_j = \Sigma'_j$.[1]

As stated, the model is never identifiable because we can always permute the $(\alpha_j, \mu_j, \Sigma_j)$ and obtain an identical distribution. What we actually care about is *identifiability up to permutation*: if $p_{\alpha,\mu,\Sigma} = p_{\alpha',\mu',\Sigma'}$ then $\alpha_j = \alpha'_{\sigma(j)}$, $\mu_j = \mu'_{\sigma(j)}$, and $\Sigma_j = \Sigma'_{\sigma(j)}$ for some permutation σ.

We have the following result:

Proposition 1.1. *As long as the orders pairs (μ_j, Σ_j) are all distinct, the parameters $(\alpha_j, \mu_j, \Sigma_j)$ are identifiable up to permutation.*

Proof. This is equivalent to showing that the functions $f_{\mu,\Sigma}(x)$ defining the pdf of a Gaussian are all linearly independent (i.e., there is no non-trivial finite combination that yields the zero function). We will start by showing this in one dimension. So, suppose for the sake of contradiction that

$$\sum_{j=1}^{m} c_j \exp(-(x-\mu_j)^2/2\sigma_j^2)/\sqrt{2\pi\sigma^2} = 0, \tag{1}$$

where the c_j are all non-zero. Then integrating (1) against the function $\exp(\lambda x)$ and using the formula for the moment generating function of a Gaussian, we obtain

$$\sum_{j=1}^{m} c_j \exp(\frac{1}{2}\sigma_j^2\lambda^2 + \mu_j\lambda) = 0. \tag{2}$$

[1] We also need to worry about the case where $k \neq k'$, but for simplicity we ignore this.

Let $\sigma_{\max} = \max_{j=1}^{m} \sigma_j$, then dividing the above equation by $\exp(\frac{1}{2}\sigma_{\max}^2\lambda^2)$ and taking $\lambda \to \infty$, we see that only those j such that $\sigma_j = \sigma_{\max}$ affect the limit. If S is the set of such indices j, we obtain

$$\sum_{j \in S} c_j \exp(\mu_j \lambda) = 0, \tag{3}$$

i.e. there is a linear relation between the functions $g_{\mu_j}(\lambda) = \exp(\mu_j\lambda)$. But this is impossible, because as long as the μ_j are distinct, the largest μ_j will always dominate the limit of the linear relation as $\lambda \to \infty$, and so we must have $c_j = 0$ for that j, a contradiction.

It remains to extend to the n-dimensional case. Suppose there was a linear relation among the PDFs of n-dimensional Gaussians with distinct parameters. Then if we project to a random 1-dimensional subspace, the corresponding marginals (which are linear functions of the n-dimensional PDFs) are also each Gaussian, and have distinct parameters with probability 1. This is again a contradiction since we already know that distinct 1-dimensional Gaussians cannot satisfy any non-trivial linear relation. □

Proposition 1.1 shows that we can recover the parameters exactly in the limit of infinite data, but it doesn't say anything about finite-sample rates. However, asymptotically, as long as the log-likelihood function is locally quadratic around the true parameters, we can use tools from asymptotic statistics to show that we approach the true parameters at a $1/\sqrt{n}$ rate.

Recovery from moments. Proposition 1.1 also leaves open the question of efficient computation. In practice we would probably use k-means or EM, but another algorithm is based on the *method of moments*. It has the virtue of being provably efficient, but is highly brittle to mis-specification.

The idea is that the first, second, and third moments give a system of equations that can be solved for the parameters (α, μ, Σ): letting $p = \sum_j \alpha_j \mathcal{N}(\mu_j, \Sigma_j)$, we have

$$\mathbb{E}_p[X] = \sum_{j=1}^{k} \alpha_j \mu_j, \tag{4}$$

$$\mathbb{E}_p[X \otimes X] = \sum_{j=1}^{k} \alpha_j (\mu_j \mu_j^\top + \Sigma_j), \tag{5}$$

$$\mathbb{E}_p[X \otimes X \otimes X] = \sum_{j=1}^{k} \alpha_j (\mu_j^{\otimes 3} + 3\,\mathrm{Sym}(\mu_j \otimes \Sigma_j)), \tag{6}$$

where $\mathrm{Sym}(X)_{i_1 i_2 i_3} = \frac{1}{6}(X_{i_1 i_2 i_3} + X_{i_1 i_3 i_2} + X_{i_2 i_1 i_3} + X_{i_2 i_3 i_1} + X_{i_3 i_1 i_2} + X_{i_3 i_2 i_1})$.

In d dimensions, this yields $d + \binom{d+1}{2} + \binom{d+2}{3} \approx d^3/6$ equations and $k(1 + d + \binom{d+1}{2}) \approx kd^2/2$ unknowns. Thus as long as $d > 3k$ we might hope that these equations have a unique (up to permutation) solution for (α, μ, Σ). As an even more special case, if we assume that the covariance matrices are all diagonal, then we only have approximately $2kd$ unknowns, and the equations have a solution whenever the μ_j are linearly independent. We can moreover find this solution via an efficient algorithm called the *tensor power method*, which is a generalization of the power method for matrices, and the rate of convergence is polynomial in k, d, and the condition number of certain matrices (and decays as $1/\sqrt{n}$).

However, this method is very brittle—it relies on exact algebraic moment relations of Gaussians, so even small departures from the assumptions (like moving from Gaussian to sub-Gaussian) will likely break the algorithm. This is one nice thing about the agnostic clustering setting—it explicitly reveals the brittleness of algorithms like the one above, and (as we shall see) shows why other algorithms such as k-means are likely to perform better in practice.

Cluster recovery. An important point is that even in this favorable setting, exact cluster recovery is impossible. This is because even if the Gaussians are well-separated, there is some small probability that a sample ends up being near the center of a different Gaussian.

To measure this quantitatively, assume for simplicity that $\Sigma_j = \sigma^2 I$ for all j (all Gaussians are isotropic with the same variance), and suppose also that the μ_j are known exactly and that we assign each point x to

the cluster that minimizes $\|x - \mu_j\|_2$.[2] Then the error in cluster recovery is exactly the probability that a sample from μ_j ends up closer to some other sample $\mu_{j'}$, which is

$$\sum_{j=1}^{k} \alpha_j \mathbb{P}_{x \sim \mathcal{N}(\mu_j, \sigma^2 I)}[\|x - \mu_j\|_2 > \|x - \mu_{j'}\|_2 \text{ for some } j' \neq j] \leq \sum_{j=1}^{k} \alpha_j \sum_{j' \neq j} \Phi(\|\mu_j - \mu_{j'}\|/\sigma) \tag{7}$$

$$\leq k\Phi(\Delta/\sigma), \tag{8}$$

where $\Delta = \min_{j' \neq j} \|\mu_j - \mu_{j'}\|_2$ and Φ is the normal CDF. As long as $\Delta \gg \sqrt{\log(k/\epsilon)}$, the cluster error will be at most ϵ. Note that the cluster error depends on a *separation condition* stipulating that the cluster centers are all sufficiently far apart. Moreover, we need greater separation if there are more total clusters (albeit at a slowly-growing rate in the Gaussian case).

1.2 Clustering Under Resilience

The mixture of Gaussians case is unsatisfying because data are unlikely to actually be Gaussian mixtures in practice, yet common algorithms like k-means still do a good job at clustering data. We therefore move to the agnostic setting, and show that we only need the distributions to be *resilient* in order to cluster successfully:

Proposition 1.2. *Suppose that $p = \sum_{j=1}^{k} \alpha_j p_j$, where each p_j is (ρ, α)-resilient with $\alpha = \min_{j=1}^{k} \alpha_j$. Then given p, it is possible to output a list of $m \leq \frac{1}{\alpha}$ "candidate means" $\hat{\mu}_1, \ldots, \hat{\mu}_m$ such that the mean of each p_j is close to one of the candidates: $\min_{l=1}^{m} \|\mathbb{E}_{p_j}[x] - \hat{\mu}_l\|_2 \leq \frac{4}{\alpha}\rho$ for all j.*

Proof. Take any decomposition $p = \sum_{j=1}^{k} \alpha'_j p'_j$ where the $p_{j'}$ are (ρ, α)-resilient and $\alpha_{j'} \geq \alpha$. Then we claim that for each j, $\mathsf{TV}(p_j, p'_{j'}) \leq 1 - \alpha$ for some j'. □

[2] This is not quite optimal, in reality we would want to assign based on $\|x - \mu_j\|_2^2/\sigma^2 + \log \alpha_j$, but we consider this simpler assignment for simplicity.