Algorithmic Statistics

Lecture: PCA and Spectral Clustering

Today, a new class of distn's and a new class of algs. to learn them.

Def: Let $\mathcal{D}$ be a class of distributions.
A mixture-of-$\mathcal{D}$ is a distribution $\sum_{i=1}^{k} w_i D_i$ where $w_i \geq 0$, $\sum_{i=1}^{k} w_i = 1$, and $D_1 \ldots D_k \in \mathcal{D}$.

Ex: Mixture of Gaussians in one dimension

Q: Under what assumptions can we learn mixtures?

Q: When can we accurately cluster samples from a mixture?

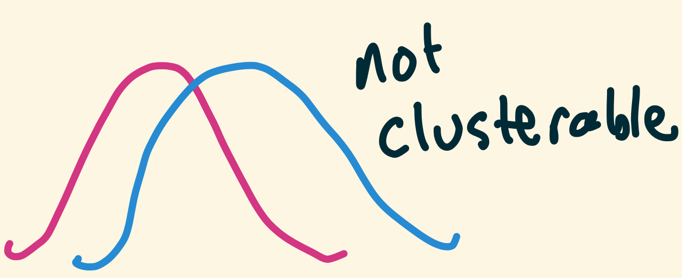
Even in one dim., can be difficult to understand, but usually algorithmically "easy".

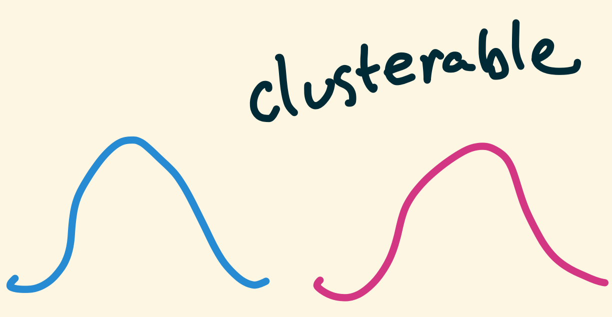
E.g. "six moments suffice" for mixture of 2 Gaussians [Kalai-Moitra-Valiant '10]

<u>Ex</u>: $\frac{1}{2} N(a,1) + \frac{1}{2} N(b,1)$. Many possible approaches:

- tournament among all possible $(\hat{a}, \hat{b})$ pairs
- use sample mean as threshold value
- k-means
- ...

works especially well if the mixture is "well-separated" / "clusterable" — for most samples, can make a good guess which component it is from.





For this simple example, clusterable iff $|a-b| \gg 1$.

k-means, Lloyd's alg

Can even analyze the "canonical" alg – Lloyd's k-means alg in one dim.

(Once we go to high dim., Lloyds still used as heuristic but will not always be right alg.)

<u>Def</u>: k-means objective: for a dataset $X_1 \dots X_n \in \mathbb{R}$, clustering $C_1 \dots C_k$ partition of $[n]$,

mean of $C_j = \frac{1}{|C_j|} \sum_{i \in C_j} X_i$

$$\text{k-means}(C_1 \dots C_k) = \sum_{j=1}^{k} \sum_{i \in C_j} |X_i - \mu_j|^2$$

Non-convex objective wrt $C_1 \dots C_k$.

Lloyd's alg – heuristic used to try to find minimizer / small-cost solutions.

Does not solve k-means in a worst-case sense – that's NP-hard.

But can hope to analyze for "nice" data – e.g. samples from a clusterable mixture.

Lloyd's alg:

1. Initialize: pick random $y_1 \dots y_k$ from $\{x_1 \dots x_n\}$ can do other initialization schemes
2. Iterate until convergence or some other termination criterion:
   a) let $C_j = \{x_i : y_j = \underset{y \in y_1 \dots y_k}{\operatorname{argmin}} |x_i - y|\}$
   b) let $y_j = \frac{1}{|C_j|} \sum_{i \in C_j} x_i$

Lemma: Spse. $x_1 \dots x_n \in \mathbb{R}$ partition into S_1, S_2 of equal size s.t. $|\mu(C_1) - \mu(C_2)| \geq \Delta$, $Var(S_1), Var(S_2) \leq 1$. W.p. $\geq 1/2$, one iter. of Lloyd's alg. finds clustering C_1, C_2 s.t.

$$\min \{ |C_1 \Delta S_1| + |C_2 \Delta S_2|, |C_1 \Delta S_2| + |C_2 \Delta S_1| \} \leq O\left(\frac{n}{\Delta^2}\right)$$

With more work, could upgrade to high probability, random samples from $\frac{1}{2} D_1 + \frac{1}{2} D_2$, larger k, ...

Proof: Condition on event that random initialization chooses one pt. from each cluster, $y_1 \in C_1$, $y_2 \in C_2$.

Chebyshev: $\Pr_{i \sim C_1} (|x_i - \mu(C_1)| > 0.1\Delta) \leq O\left(\frac{1}{\Delta^2}\right)$.

Triangle inequality: all but $O(1/\Delta^2)$ pts in C_1 are closer to y_1 than to y_2, similarly for C_2, y_2, with const. probability. $\square$

Now $d > 1$ dimensions. How far apart do D_1, D_2 need to be for similar naive alg to work? (Normalization: $\mathrm{Cov}(D_1), \mathrm{Cov}(D_2) \preceq I$).

Success relied on

$$\underbrace{\mathbb{E}_{i \sim C_1} \| x_i - \mu(C_1) \|}_{} < \| \mu(C_1) - \mu(C_2) \|$$

This is like $\sqrt{d}$ in d dimensions!

$\Rightarrow$ Naive clustering algs require $\| \mu(D_1) - \mu(D_2) \| > \sqrt{d}$.

Idea: use PCA to reduce dimension first!

Interlude: Spiked Matrix Models & Best Low-Rank Approximation

Suppose you observe a matrix $M \in \mathbb{R}^{m \times n}$ form

$$M = \overset{\geq 0}{\lambda} \cdot \underbrace{uv^T}_{\text{rank } 1,\text{ or more generally rank } r} + \underset{\text{random}}{W}$$

Different distn's of W are studied - for now let $W_{ij} \sim N(0,1)$, normalize s.t. $\|u\| = \|v\| = 1$.

Aside: it's a graphical model inference prob. to estimate u, v or uv^T given M.

Naive estimator of uv^T - $\frac{1}{\lambda} M$.

$$\mathbb{E} \frac{1}{\lambda} M = uv^T,$$

$$\mathbb{E} \| uv^T - \frac{1}{\lambda} M \|_F^2 = \frac{1}{\lambda^2} mn$$

would need $\lambda > \sqrt{mn}$ to be nontrivial

Idea: use best rank-one approximation of $\frac{1}{\lambda} M$.

Let X be best rank-one approx, ie

$$X = \underset{Y \text{ rank } 1}{\arg\min} \left\| Y - \frac{1}{\lambda} M \right\|_F^2$$ — not convex but solve via SVD.

<u>Theorem:</u> $\mathbb{E} \|X - uv^T\|_F^2 \leq O\left(\frac{m+n}{\lambda^2}\right)$

To prove thm., use that $\mathbb{E}\|W\|_{op}^2 \leq O(n+m)$
(provable up to logs via matrix Bernstein.)

Denote $E = X - uv^T$ = error matrix.

<u>Claim 1:</u> $\|E\|_F^2 \leq 2 \langle E, \frac{1}{\lambda} W \rangle$

<u>Pf:</u> $0 \leq \|uv^T - \frac{1}{\lambda} M\|_F^2 - \|X - \frac{1}{\lambda} M\|_F^2$

$$= \left\|\frac{1}{\lambda} W\right\|_F^2 - \left\|E - \frac{1}{\lambda} W\right\|_F^2$$

$$= \left\|\frac{1}{\lambda} W\right\|_F^2 - \|E\|_F^2 + 2\langle E, \frac{1}{\lambda} W\rangle - \left\|\frac{1}{\lambda} W\right\|_F^2$$

$$= -\|E\|_F^2 + 2 \langle E, \frac{1}{\lambda} W \rangle \qquad \square$$

<u>Claim 2:</u> $\underbrace{\|E\|_*}_{\text{nuclear norm}} \leq \sqrt{2} \|E\|_F$

Pf: $\mathrm{Rank}(E) \leq 2$. $\blacksquare$

Claim 3: $\|E\|_F^2 \leq O\left(\frac{1}{\lambda^2}\|W\|_{op}^2\right)$

Pf: by Claims 1, 2, and Hölder,

$$\|E\|_F^2 \leq 2\sqrt{2} \cdot \frac{1}{\lambda} \cdot \|W\| \cdot \|E\|_F.$$

Cancel $\|E\|_F$ from both sides and square. $\blacksquare$

Theorem follows from claims 1-3 + $\mathbb{E}\|W\|^2 \leq O(n+m)$.

Same proof for rank r gives error $O\left(\frac{r(m+n)}{\lambda^2}\right)$.

(End of Interlude)

Back to clustering/learning mixtures.

Concretely, consider $\frac{1}{2}N(\mu_1, I) + \frac{1}{2}N(\mu_2, I)$.

From before: if $\|\mu_1 - \mu_2\| > \sqrt{d}$, naive algs/proofs don't work.

If we knew the direction $\mu_1 - \mu_2$, could project samples in that direction, turn it into a 1-dim. problem.

Direct calculation shows:

$$\mathrm{Cov}\left(\tfrac{1}{2}N(\mu_1, I) + \tfrac{1}{2}N(\mu_2, I)\right)$$
$$= I + \tfrac{1}{4}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

So should be possible to find $\mu_1 - \mu_2$ from samples (approximately).

Given samples $X_1 \ldots X_n$, estimate $\mu_1 - \mu_2$

via best rank-1 approx. to

$$M = \frac{1}{n} \sum (x_i - \hat{\mu})(x_i - \hat{\mu})^T - I$$

Can show via e.g. Matrix Bernstein that

$$M = \frac{1}{4}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T + W, \quad \text{where}$$

$\mathbb{E}\|W\|_{op}^2 \leq \tilde{O}\left(\sqrt{\frac{d}{n}}\right)$ if $n \geq d$.

Same argument as before: get estimator $X = xx^T$ s.t. $\mathbb{E}\|X - (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T\|_F^2 \leq \tilde{O}\left(\sqrt{\frac{d}{n}}\right)$.

$$\|xx^T - (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T\|_F^2$$

$$= \|x\|_2^4 - 2\langle x, \mu_1 - \mu_2 \rangle^2 + \|\mu_1 - \mu_2\|_2^4$$

So, $\|x\|_2^2 \, \|\mu_1 - \mu_2\|_2^2 \underset{(AM-GM)}{\leq} \langle x, \mu_1 - \mu_2 \rangle^2 + \tilde{O}\left(\sqrt{\frac{d}{n}}\right)$

$$\Rightarrow \frac{|\langle x, \mu_1 - \mu_2 \rangle|}{\|x\| \cdot \|\mu_1 - \mu_2\|} \geq 1 - \frac{\tilde{O}\left(\sqrt{\frac{d}{n}}\right)}{\|\mu_1 - \mu_2\|^2}$$

Exercise (tedious but easy): this is a good-enough approx. of direction $\mu_1 - \mu_2$ to reduce to 1-dim case. Hence, $\|\mu_1 - \mu_2\| \gg 1$ sufficient.

<u>Extension</u>: Mixture of k $\rightarrow$ reduce to $O(k)$-dim subspace $\text{span}(\{\mu_i\})$.

Pairwise distances $\geq \Omega(\sqrt{k})$ suffice.